Dual-Stream Multiple Instance Learning for Depression Detection With Facial Expression Videos

Zixuan Shangguan, Zhenyu Liu, *Member, IEEE*, Gang Li, Qiongqiong Chen, Zhijie Ding, and Bin Hu, *Senior Member, IEEE*

***Abstract*— Depression is a common mental illness which has brought great harm to the individuals. With recent evidence that many objective physiological signals are associated with depression, automated detection of depression is urgent and important for the growing concern of mental illness. We investigate the problem of classifying depression by facial expressions, which may aid in online diagnosis and rehabilitation engineering of depression. In this work, We propose a weakly supervised learning approach employing multiple instance learning (MIL) on 150 videos data from 75 depressed and 75 healthy subjects. In addition, we present a novel MIL dual-stream aggregator that considers both the instance-level and the bag-level in order to emphasize the information with symptoms. Specifically, our method named ADDMIL uses max-pooling at the instance level to capture symptom information and further integrates the contribution of each instance at the bag level using attention weights. Our method achieves 74.7% accuracy and 74.5% recall on the collected dataset, which not only improves 10.1% accuracy and 9.8% recall over the baseline but also exceeds the best accuracy result of MIL-based method by 2.1%. Our work achieves results that are comparable to the state-of-the-art methods and demonstrates that multiple instance learning has great potential for depression classification. We present for the first time a weakly supervised learning approach in the detection of depression through raw facial expressions, which may provide a new framework for other psychiatric disorders detection methods.**

Manuscript received 13 May 2022; revised 7 August 2022; accepted 30 August 2022. Date of publication 6 September 2022; date of current version 1 February 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFA0706200; in part by the National Natural Science Foundation of China under Grant 61632014, Grant 61627808, Grant 61802159, and Grant 61802158; and in part by the Fundamental Research Funds for Central Universities under Grant lzujbky-2019-26 and Grant lzujbky-2021-kb26. *(Corresponding authors: Zhenyu Liu; Bin Hu.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Medical Ethics Committee of Tianshui Third People's Hospital.

Zixuan Shangguan, Zhenyu Liu, and Bin Hu are with the Gansu Provincial Key Laboratory of Wearable Computing, Lanzhou University, Lanzhou, Gansu 730000, China (e-mail: shanggzx20@lzu.edu.cn; liuzhenyu@lzu.edu.cn; bh@lzu.edu.cn).

Gang Li and Zhijie Ding are with the Tianshui Third People's Hospital, Tianshui 741000, China (e-mail: lg13689402968@163.com; dingzj_2004@163.com).

Qiongqiong Chen is with the Second Provincial People's Hospital of Gansu, Northwest Minzu University, Lanzhou 730030, China (e-mail: cqq-xingcai@163.com).

Digital Object Identifier 10.1109/TNSRE.2022.3204757

***Index Terms*— Depression detection, multiple instance learning, weakly supervised learning, facial expression.**

I. Introduction

MAJOR Depressive Disorder (MDD), also referred to as depression, is a globally common mental disorder and causes physical and mental health damage to hundreds of thousands. Unlike the usual mood swings and transient emotional responses to the challenges of daily life, depression will bring pervasive low mood, lack of confidence, and loss of pleasure or interest in activities for most of the day. Moreover, depression could increase the risks of diabetes, heart disease, cancer [1], and in serious cases, it can lead to suicide.

At present, there are effective psychological and pharmacologic treatments like antidepressants, Dialectical Behavior Therapy (DBT), and Cognitive Behavior Therapy (CBT). However, due to the lack of medical resources, trained healthcare workers, and social prejudice against mental disorders, people with depression are often not properly diagnosed and treated. More specifically, across all countries at different income levels, people with depression are often underdiagnosed, and even non-depressed people are often misdiagnosed and prescribed antidepressants [2]. Given the high incidence of depression and the lack of appropriate treatments for large populations, methods that rely solely on subjective assessment and diagnosis are no longer able to meet current medical needs. Therefore, automatic detection of depression will be very helpful and necessary for the diagnosis of this mental disorder.

Some clinical literatures have shown that depressive states can be expressed from facial expressions [3], [4]. Compared with the healthy subjects, the facial expressions of depressed subjects are usually neutral or sad expressions, which are characterized by frowning, drooping eyes, and looking tired or worried [3], [4]. Several studies have attempted to automatically detect depression through facial information. For example, Zhu *et al.* [5] proposed a framework with a two steams manner aiming to capture the appearance and dynamics of a subject for depression assessment. Ding *et al.* [6] proposed a deep learning framework based on facial expression recognition to classify depression and remission or response to treatment. In [7], Melo *et al.* introduced a 3D convolutional neural network (CNN) with different temporal depths and

receptive field sizes to better produce spatio-temporal features from facial videos.

Most of the state-of-the-art works focusing on detecting depression disorder from facial expression are supervised learning approaches, which identify each instance (frame or clip) in a video with a specified label during training. Typically, a supervised model classifier requires annotated data that clearly indicates which instances contain the desired video event and which instances do not. Although subjects' videos contained a large number of natural facial expressions and dynamic changes, we were unable to precisely determine which instances in the videos reflected information related to depression level. In the current methods, the final prediction result of the subject is achieved by averaging the scores of each instance, which means that each instance contributes equally to the outcome. Due to the powerful fitting of deep learning, some instances that do not show obvious symptoms may negatively affect the model performance.

To address these challenges, we collected videos of facial expressions in patients with confirmed depression watching negative video stimuli, which may exacerbate the corresponding symptom expression. We treat the description of such data as a "weakly labeled", which provides information on the presence or absence of depression but does not specify detailed information such as the precise times in the video that indicated the disorders, or the regions identified for the duration of these disorders. We use weakly labeled data to identify subjects with high depressive symptoms and effects.

Therefore, the situation of detecting depression by facial expressions conforms to the rules of multiple instance learning. For convenience, we consider the facial video of each subject as a bag containing many instances (frame or clip). Positive and negative bags are used to represent bags with and without highly depressive symptoms, respectively. Herein, we propose for the first time the use of weakly supervised learning to accurately detect depression by employing a novel attention based dual-stream deep multiple instance learning (ADDMIL) method. Our proposed ADDMIL model combines a score of max-pooling instance, which we refer to top instance, and an attention score for each instance by measuring the distance from each instance to the top instance. The use of max-pooling ensures that we can capture the maximum extent of symptoms and states of mental disorder, and the use of attention weights allows us to synthesize the overall information of the bag. To evaluate the classification accuracy of our approach, we compared our model to several recently MIL models on our depression facial dataset. The results show that our method outperforms other recent MIL models by at least 2.1% in classification accuracy. Compared to state-of-the-art supervised learning methods, our method still achieves better performance in classification accuracy.

The rest of this paper is arranged as follows. In Section II, we introduce the related work for depression detection; In Section III, we present some preliminaries and details of our proposed model; In Section IV we provide the experimental results; We discuss and conclude our study in Section V and Section VI.

II. Related Work

This section briefly presents related works from three aspects: 1) traditional hand-engineered feature methods, 2) deep learning methods and 3) weak supervision and multiple instance learning method.

A. Traditional Hand-Engineered Method

Many automatic depression recognition studies used hand-crafted features for depression facial experessions representations. Recently, the Audio-Visual Emotion Challenge 2013 [8] and 2014 [9] (AVEC 2013 and AVEC 2014) provided depression video dataset and hand-engineered feature methods have been proposed in these competitions. The baseline of AVEC2013 used the Local Phase Quantization (LPQ) [10] as a facial descriptor and then evaluated depression level by the Support Vector Regression (SVR) [11]. Meng *et al.* [12] used the motion history histogram (MHH) [13] to model the motion cue during videos. Two different descriptors, the Space-Time Interest Points (STIP) [14] and Pyramid of Histogram of Gradients (PHOG) [15], were employed in [16] for depression analysis. Wen *et al.* [17] proposed to encode temporal information based on LPQ from Three Orthogonal Planes (LPQ-TOP) features from sub-volumes of facial region. In the AVEC2014 competition, Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [18] was the baseline facial features that combine temporal and spatial information based on Gabor filtering. Jan *et al.* [19] introduced a 1D MHH calculated on feature representations of local descriptors. Kaya *et al.* [20] combined LGBP-TOP and LPQ features as the video representation to predict depression level.

Methods based on hand-engineered features are more interpretable and focus on spatiotemporal information. However, hand-engineered features require a lot of prior knowledge and lack high-level semantic features, which can lead to feature redundancy and time-consuming.

B. Deep Learning Methods

More recently, several deep learning-based methods have been applied to address depression detection. For example, Jazaery *et al.* [21] used the convolutional 3D network (C3D) [22] to capture the facial spatiotemporal features of two different scales. Then a Recurrent Neural Network (RNN) [23] was employed to learn the features of consecutive clips. Melo *et al.* [24] proposed a MDN module composed of two blocks: maximization block and difference block, which can capture smooth facial variations and encode sudden transitions of facial structures respectively. Moreover, Melo *et al.* [25] considered the relationship between facial expression and depression levels labels, and proposed to use distribution learning to reduce the impact of noisy labels, however, they did not explore the relationship between affective states and labels. In [26], the authors introduced a multimodal deep learning network with appearance and dynamics of facial information. Instead of a simple decision level fusion, they introduced a chained-fusion mechanism for modeling correlation and complementarity information to improve depression detection.

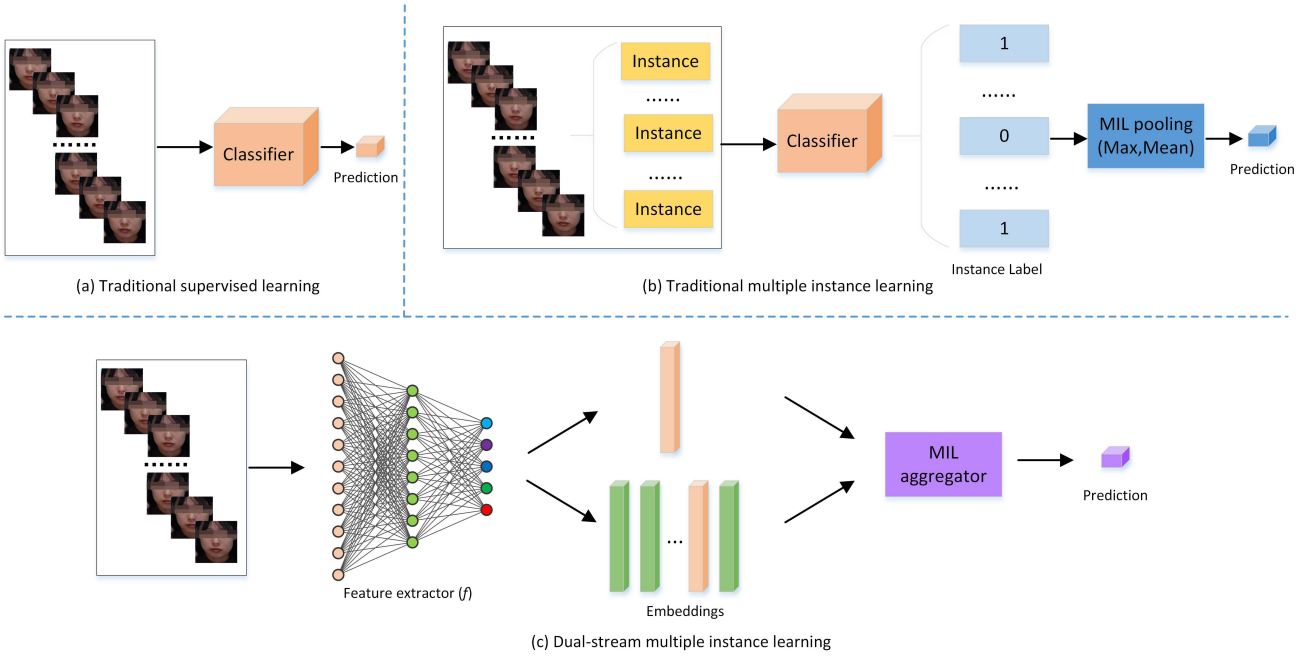

Fig. 1. The illustration of the proposed dual-stream deep multiple instance learning compared with traditional learning paradigms: (a) overview of previous supervised learning in depression detection; (b) overview of traditional MIL in depression detection: after classifying each instance individually, traditional MIL pooling e.g., max pooling and mean pooling are used; (c) the ADDMIL method proposed by us, after passing through the feature extractor *f*, respectively inputs the embeddings required by the dual-stream MIL aggregator and obtains the prediction results.

The methods based on deep learning mainly consider using 2D CNN to extract spatial information or using 3D CNN to extract spatiotemporal information, which contains high-level semantic features to make the model achieve better performance. Although both hand-engineered feature based and deep learning based methods have contributed significantly to depression detection, they both assume supervised learning methods using strongly labeled data. However, in practical situations we cannot guarantee that such an assumption holds, so we use a weak supervision approach to explore the benefits of solving this problem.

C. Weak Supervision and Multiple Instance Learning

Multiple instance learning (MIL) is a form of weakly supervised learning [27]. In a general MIL task, labels are invisible for individual instances but are visible to a collection of instances, which called bag. In classification tasks with weakly labeled data, using supervised learning classifiers can introduce label noise. The performance of MIL adaptation in this context can provide better solutions due to supervised learning. At present, many powerful algorithms of MIL manifest and perform at four levels: instance-level [28], bag-level [29], embedding-based, and some joint approach incorporating other methods such as attention mechanisms [30], [31]. Moreover, MIL has been widely used in various domains such as tumor detection [32], object detection [33], action localization [34], image classification [35] and latest COVID-19 screening [31], [36].

In literature, MIL has been applied in the affective computing domain due to its superior structure and backbone. In the field of facial expression recognition (FER) [37], [38], MIL has been widely used and achieved promising performance [39], [40]. Xie *et al.* [41] proposed a MIL method called MMED for early expression detection (EED) for the first time, whose task is to detect an expression as soon as possible. Wu *et al.* [42] employed MIL to predict the engagement level of students watching the education video in the EmotiW Challenge 2020 [43]. MIL also significantly improved the performance of depression detection. Salekin *et al.* [44] introduced that although the voice speech data related to depression diagnosis had labels, it did not provide other detailed information, such as the voice interval or segment to show depression states, and proposed a weakly supervised learning framework for MIL. Similarly, Ren *et al.* [45] proposed a MIL approach for bipolar disorder diagnosis by weakly labelled speech data. More recently, in the facial landmark expression-based approach from [46], the authors utilized both feature manipulation and MIL to handle the coarse-grained labels contained in the video clips and got a state-of-the-art performance. However, using the video data of landmark would lose smooth details in facial expression, and the MIL used in this approach only contained the instance level method which was considered inferior to the other level MIL performance [30], [47].

Inspired by recent works, we treated the collected depression data as weakly labeled, which only indicated the presence or absence of depression without specifying specific information. This description is more in line with the data of depression detection and is applicable to the rules of multi-instance learning, so we propose a novel MIL framework based on both attention and max-pooling instance for depression detection. To the best of our knowledge, our work is the first time to employ MIL technology for depression detection via raw facial expression. Extensive experiments on our depression databases show that our method performs better or matches the other state-of-the-art methods.

III. Method

This section presents our necessary notations and objectives for the task of depression detection, including the formulation of MIL and details of the relative module in our work.

A. Preliminaries

The MIL model receives a set of n labeled sample pairs $\{(S_i, Y_i)\}_{i=1}^{n}$ drawn from the joint distribution defined by $S \times Y$, wherein S_i is the whole video session and Y_i is $\{0, 1\}$ for binary classification corresponding to the session. Also, $S_i = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_M\}$ is a bag of M instances (i.e., frames in whole video session) and each instance s_m has a label $y_m \in \{0, 1\}$. Furthermore, in our task, these instances can be supposed to be positive or negative and instances in a bag may not all positive or negative. That means not all instances in a video session show depression symptom, as other instances may be noise and not helpful for learning.

In literature, traditional MIL studies must satisfy the following constraints: S_i is a positive bag if there exists at least one positive instance; otherwise, if a bag S_i is negative, then all corresponding instances should be negative. The label of bag is given by

$$Y = \begin{cases} 0, & \text{iff } \sum_m y_m = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

In our work, considering the situation for depression detection from facial expression during a video session is more complicated because two groups of bags may contain both negative and positive instances, this assumption may not hold strictly. Therefore, on the basis of considering the instance level, we also consider to incorporate implicit instance weights in bag level, which can be seen as a looser version of the constraint.

Algorithm 1 ADDMIL Algorithm

Input: parameters $\theta_f, \theta_{g_1}, \theta_{g_2}, W_m, W_s$, epoch T
Output: $\theta_f, \theta_{g_1}, \theta_{g_2}, W_m, W_s$
initialize parameters $\theta_f, \theta_{g_1}, \theta_{g_2}, W_m, W_s$
for $i = 1, 2, \ldots, T$ **do**
  preprocess 2D facial video frames $[s]_{i=1}^{n}$
  obtain features:$h = f(s)$
  produce max-pooling prediction $Z_1 = g_1(s)$
  obtain attention weights β by Eq. (3) and Eq. (4)
  combine instance representation $e = \sum_{i=0}^{N-1} \beta v_i$
  produce bag prediction $Z_2 = g_2(h_1, \ldots, h_n) = W_s e$
  obtain final prediction $Z = \frac{Z_1 + Z_2}{2}$
  update parameters $\theta_f, \theta_{g_1}, \theta_{g_2}, W_m, W_s$,
end for

B. ADDMIL for Depression Detection

As illustrated in Figure 1, our approach takes a video frame of unlabeled instances and transforms them into semantic representations. Compared with traditional multiple instance learning, ADDMIL first transforms the instance j from bag i in the video session bag into the corresponding embedding $h_{ij} = f(s_{ij}) \in \mathbb{R}^L$ by a feature extractor model f. Following, we use an aggregation function g that employs permutation invariant to obtain the final prediction $\hat{Z}_i = g(h_{i1}, \ldots, h_{iM})$.

Our most critical innovation lies in how to design a novel aggregate function g to find information about high symptoms from weakly labeled video data. Previous MIL aggregators have focused on instance-level and bag-level methods, usually in the form of max-pooling or mean-pooling [48], [49]. We employ the max-pooling operation at the instance level to identify instances indicating high depression symptoms. Also, we use non-local operations to model dependencies between instances and bag to obtain contextual information. Similar to the self-attention method, the only consider the relationship of the top instance to other instances which strengthen the depression information in bag level. We now describe the detailed component of ADDMIL.

C. Dual-Stream MIL Aggregator of ADDMIL

Recently, many studies have attempted to use deep networks to incorporate the MIL framework [50]. Among them, Ilse *et al.* [30] proposed to use an attention mechanism to integrate the contribution of instances into bag embedding. Moreover, Rymarczyk *et al.* [47] formed a new pooling aggregator by combining attention mechanisms and self-attention [51], which not only integrated the overall information of instances but also took into account the interdependencies among instances. Li *et al.* [32] consider the use of non-local operations to account for the aggregation process between the bag and the instances embedding. Inspired by these, ADDMIL uses a two-stream structure to comprehensively consider the construction of aggregator pooling layers. The detailed algorithm is presented in Algorithm 1.

For brevity, we denote $B = \{s_1, \ldots, s_n\}$ as a bag of frames and f as a feature extractor. Each instance becomes the corresponding embedding $h_i = f(s_i)$ with $h_i \in \mathbb{R}^L$ through f. In first stream we employ max-pooling at the instance level to get the instance with the highest score, which is called the top instance. It follows that

$$\begin{aligned} Z_1 &= g_1(f(s_1), \ldots, f(s_n)) \\ &= \max\{W_m h_0, \ldots, W_m h_{N-1}\} \end{aligned} \tag{2}$$

where W_m is a weight vector. With the max-pooling stream we get the highest scoring instance h_m (top instance), which most likely represents a depressive state.

We obtain the embedding of the bag by aggregating the embeddings of the related instances in the second stream. Specifically, we transform the embeddings of each instance into query q_i and information v_i, with $q_i \in \mathbb{R}^L$ and $v_i \in \mathbb{R}^L$, and calculate:

$$q_i = W_q h_i, \quad v_i = W_v h_i \tag{3}$$

where W_q and W_v are both a weight matrix. Likewise, we define the distance measurement β from all instances to the top instance here:

$$\beta(h_i, h_m) = \frac{\exp(o_{i,m})}{\sum_{k=0}^{N-1} \exp(o_{k,m})}, \text{ where } o_{(i,j)} = \langle q_i, q_j \rangle \tag{4}$$

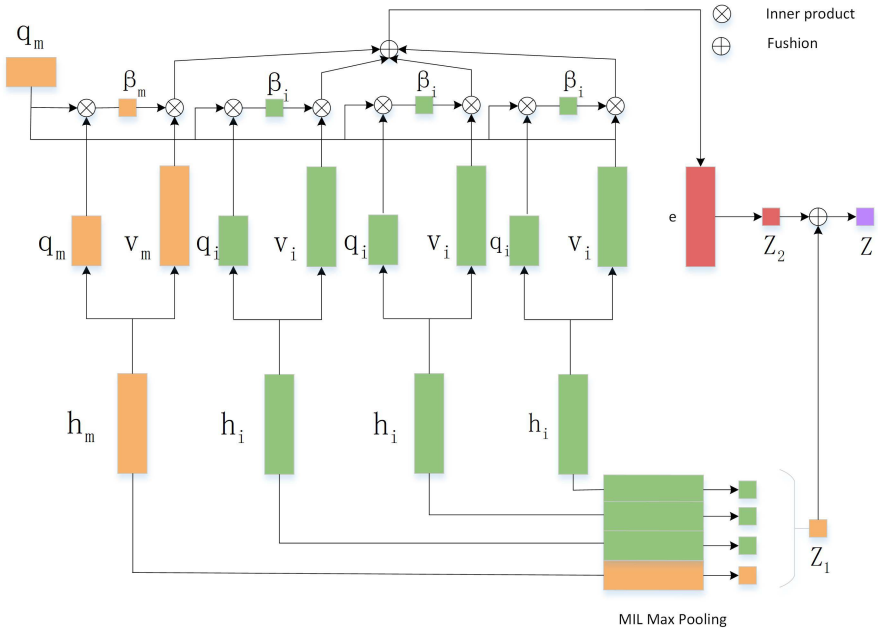

Fig. 2. The MIL aggregator for ADDMIL consists of two streams. One of the streams is to pick the top instance with the highest score at the instance level via max-pooling. Another stream is by taking the distance of other instances to the top instance as a weight and then getting a bag-level embedding and getting its score. Among them, the green column represents the embedding of the top instance and the yellow column represents the embedding of other instances. The final score is the average of the two streams.

with $\langle \cdot, \cdot \rangle$ denotes inner product operation of vector. By using the distance measurement β as the weight, the embedding of the bag is defined as:

$$e = \sum_{i=0}^{N-1} \beta\left(h_i, h_m\right) v_i \tag{5}$$

The output of bag score is defined as:

$$\begin{aligned} Z_2 &= g_2\left(f\left(s_1\right), \ldots, f\left(s_n\right)\right) \\ &= W_s \sum_{i=0}^{N-1} \beta\left(h_i, h_m\right) v_i = W_s e \end{aligned} \tag{6}$$

with W_s is a weight vector. Different from the self-supervision mechanism, in order to reduce the influence of noisy nodes, we only consider the distance measurement between key nodes to other nodes. Moreover, the weights are not obtained from a two-layer neural network like an attention mechanism, but are given by measuring the similarity of an arbitrary instance to a top instance through an inner product operation, which means that the instance with higher similarity to the top instance will have a larger attention weight. We ensure that the sum of the weights is 1 to be invariant to the size of a bag through the operation of Equation 4. The score for the final bag is given by:

$$\begin{aligned} Z &= \frac{1}{2}\left(g_1\left(f\left(s_1\right), \ldots, f\left(s_n\right)\right) + g_2\left(f\left(s_1\right), \ldots, f\left(s_n\right)\right)\right) \\ &= \frac{1}{2}\left(W_m h_m + W_s e\right) \end{aligned} \tag{7}$$

The bag's score Z represents the probability of depression, when $Z > \tau$, the final label of the bag is $Y = 1$ (Depressed), otherwise the final label of the bag is $Y = 0$ (Healthy). In this paper, we set the threshold τ to 0.5.

From a technical standpoint, different weights are assigned to instances according to their similarity to top instance, thus implicitly satisfying the feature selection for the information vector v_i. The synergistic combination of max-pooling and attention-based pooling enables the bag to be more informative. The architecture of the DSMIL aggregator is showed in Figure 2.

IV. Experiments

A. Dataset

The current state-of-the-art methods have been applied to the AVEC2013 and AVEC2014 datasets, but are not disclosed due to privacy concerns. In order to obtain visual depression data with high quality, our datasets are collected from Tianshui Third People's Hospital, a professional psychiatric hospitals. The subjects who participated in the experiment went through a professional psychiatrist and questionnaires such as the Mini-International Neuropsychiatric Interview (MINI) and the Patient Health Questionnaire-9 (PHQ-9), where the doctor's diagnosis and PHQ-9 was considered the main grouping criterion. Specifically, our subjects were divided into a healthy control group and a depressed patient group according to the doctor's diagnostic opinion and whether the PHQ score was less than 5. Moreover, one-to-one matching operation is performed on the data according to age, gender and educational background. In the matching process, the age difference is required to be within three years; the gender is uniform; the educational level difference is within 3 (the educational level of the subjects is determined by the educational level, illiterate is 0, primary school is 6, junior high school is 9,

TABLE I
STATISTICAL CHARACTERISTICS OF THE SUBJECTS: MEAN AND (STANDARD DEVIATION)

Subject	Number	Age	Education	PHQ-9
Depressed	75	35.92(10.12)	11.5(3.61)	18.31(8.40)
Healthy	75	35.61(9.97)	12.3(3.97)	2.78(0.88)

high school is 12, and university is 16, postgraduate and above is 19).

For collecting rigorous data, our experimental collection was completed in a quiet, undisturbed, soundproof room. At the same time, we used the Logitech CC2000e camera with pixels up to 1920 × 1080, 30 frames per second, and 10x lossless zoom to capture clearly face varieties. In the experimental task, subjects were asked to watch a negative video clip from China's standard Emotional Video Stimuli materials (CEVS) about two minutes in length. The entire experimental process is reviewed and approved by the ethics committee, without the use of drugs or other means and the subject's private data is also protected. During the video capture process, the subjects were asked to face the camera while maintaining a distance of about 50 cm from the camera to ensure clear facial images were captured. Excluding the data corrupted by factors such as facial occlusion and improper operation of the test personnel, we finally selected 150 video clips from 75 depressed patients and 75 healthy subjects. Detailed subject statistics are presented in the Table I.

B. Experimental Setup

Accordingly, our dataset is divided into training, validation and test sets according to the subject ID with a ratio of 8:1:1, and all experiments are subject to 10-fold cross-validation. Each face region in the video is cropped and aligned by machine learning toolkit Dlib [52] and resized to 224 × 224. To reduce the temporal redundancy in videos, we empirically extract one frame out of every 10 frames. The total number of video frames is about 45,000 frames, and each bag contains about 300 instances.

Inspired by Kim *et al.* [53], ADDMIL uses AlexNet [54] as a facial feature extractor f and is pre-trained on the AffectNet dataset [55], which consists of about 440,000 over a million images for automated facial expression recognition. The Alexnet employ Adaptive moment estimation (ADAM) with learning rate of 1e-4 and the mini-batch size is set to 256. We reduce the learning rate by 0.8 times every 10k iterations. Process of the pre-training takes about 100,000 iterations. In the fine-tuning stage, the ADAM is adopted with initial learning rate of 0.002 and reduce the learning rate by a cosine annealing scheme [56]. Eps and weight decay of the ADAM optimizer are set to the values of 0.0005 and 1e-8. The process in our work is trained for 30 epochs with mini-batch size of 1 (bag). The proposed model training in our experiment is implemented in Pytorch. We report the metrics of accuracy, recall, precision and F1 for performance evaluation of the task of depression detection.

TABLE II
EVALUATION OF THE SEVERAL MIL METHODS ON OUR DATASET

Method	Accuracy	Precision	Recall	F1
Max-pooling [48]	0.680	0.684	0.712	0.673
Mean-pooling [48]	0.686	0.613	0.613	0.630
Attention based [30]	0.726	0.665	0.663	0.652
ADDMIL	**0.747**	**0.724**	**0.745**	**0.723**

C. Experimental Results

1) Performance of MIL Models: To explore the effectiveness and feasibility of depression detection in the multiple instance learning domain, we first introduce several MIL methods applied in the detection of depression. We consider to use three MIL methods here including max-pooling based [48], mean-pooling based [48] and attention-based [30] pooling model to compare with our method ADDMIL.

In particular, methods based on max-pooling and mean-pooling comply with the traditional assumptions of MIL on instance level learning, where the final subject-level predictions are derived from the top instance and the mean of all instance in each bag, respectively. Method of Attention based pooling uses the attention mechanism to weight the embedding of each instance on bag level learning. In contrast, ADDMIL takes into account both instance and bag-level learning by using dual-stream MIL. It is worth noting that all MIL methods use the same feature extractor. The network structure of extracting the top instance stream in ADDMIL is the same as that of the traditional MIL pooling, while the other stream is different from the network architecture based on the attention pooling due to the different method design.

As show in Table II, our proposed ADDMIL achieves the highest accuracy overall performance with accuracy of 74.7%, which consistently outperform the best results of method based on MIL by 2.1% in terms of accuracy. Method with max-pooling reports the lowest performance of all MIL methods, which is expected since it only takes account into the top instances from 2D visual frames in the bag to predict. Interestingly, the max-pooling method still has a second highest recall that is comparable to ours among these methods. The method based on mean-pooling considers all instances, but introduces too much useless information to each instance with the same weight, which makes the final result unsatisfactory. Moreover, though pure attention based pooling improves significantly over traditional methods at the instance level, our proposed ADDMIL achieves best performance in terms of accuracy, recall, precision and F1, indicating that the effectiveness of the dual-stream MIL structure.

2) Comparison of Previous State-of-the-Art Models: We compare against several state-of-the-art depression detection methods to evaluate our proposed approach, and the results are shown in Table III. Noting that many state-of-the-art methods are used as regression problems to assess depression levels, for fair comparison, we re-applied these methods to the depression classification problem by changing the loss function and method structure. For example, we employ SVM to replace the classification layer of some traditional methods or get the classification results by modifying the number of units after

TABLE III
EVALUATION OF STATE-OF-THE-ART METHODS ON OUR DATASET

Method	Accuracy	Precision	Recall	F1
LGBP-TOP+ELM [9] (2014)	0.646	0.652	0.647	0.641
MHH-LBP+SVM [12] (2013)	0.633	0.634	0.675	0.653
LPQ-TOP+SVM [17] (2015)	0.645	0.631	0.713	0.674
C3D [21] (2018)	0.666	0.715	0.612	0.625
MSN [7] (2020)	0.734	0.712	**0.763**	**0.734**
Resnet50 [25] (2019)	0.713	0.714	0.691	0.693
MDN [24] (2021)	0.735	**0.743**	0.716	0.732
CNN+Attention [57] (2021)	0.713	0.670	0.705	0.683
Ours (ADDMIL)	**0.747**	0.724	0.745	0.723

the fully connected layer of the deep model. In general, five of them are deep learning methods, and the other three are based on handcrafted features.

In [9], the baseline method employed the LGPB-TOP descriptor as the hand-crafted features. Compared with the baseline, our method improves the accuracy by 10.1%. In Wen *et al.* [17], considered using the LPQ-TOP method to extract the dynamic information of the video, and performed dimensionality reduction on the obtained features. From the results, it can be seen that methods based on handcrafted features often perform worse than deep learning methods, which may be limited by the fact that handcrafted features are based on a large amount of prior knowledge and lose many subtle raw information.

In [21], a C3d network was used to extract spatiotemporal features in videos, and then an RNN was used to combine these features. In [7], the method of Melo *et al.* combined different scales of convolution filters in their 3D convolution layers to obtain richer spatiotemporal information, thus achieving higher performance. On the other hand, 2D deep networks have also achieved promising performance. For example, in the work of [24], the author used two blocks to comprehensively consider the subtle and sudden changes of facial expressions between frames. In [57], the authors introduced a Local-Global Attention combined with local patches and global patterns from facial information.

Our proposed method achieves the highest accuracy and the second highest recall, where recall is particularly important in the depression detection domain due to the reduction of false negatives that can lead to serious consequences including patient suicide.

3) Qualitative Results: In Figure 3, we introduce the qualitative results of our attention weights. In order to reduce the impact of the number of instances in different bags, we divide the whole bag into 10 segments for analysis. In general, the healthy subjects, as the negative bags, have a low attentional weight in each segment, while the depression subjects, as the positive bags, show a high attentional weight in a certain segment. This is expected since healthy subjects do not have sufficient information to detect depression throughout the video, and depressed subjects pay high attention in a segment because our method encourages the similarity of the top instance to the other instances.

Interestingly, we can find that segment with significantly high attentional weights mainly focus on the second half of

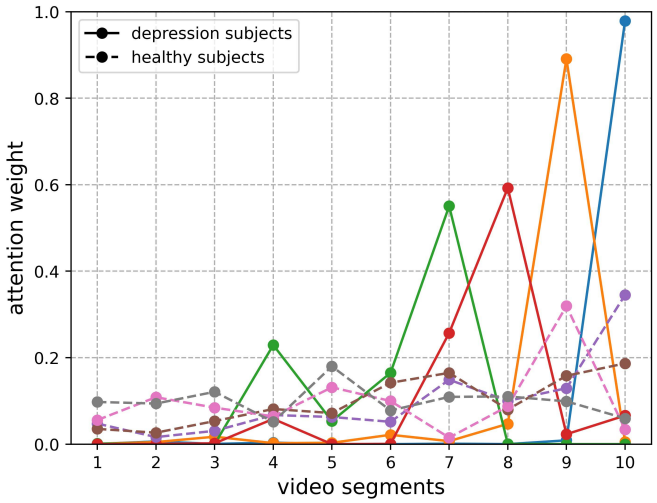

Fig. 3. The video in the figure is divided into 10 segments, and the attention weight of each segment is composed of the sum of the instances it contains. The solid line in the figure represents the depression subjects, and the dashed line represents the healthy subjects.

viewing the stimulus material. Similarly, the climax part of the negative video stimulation material we intercepted is also at the end of the video, which may indicate that sustained high-intensity negative stimuli can induce depression symptom in both healthy and depressed subjects.

4) The Effect of Dual-Stream MIL Aggregator on Learning: We consider the compositional settings of the dual-stream aggregator to evaluate the effectiveness of our proposed model. Specifically, we design the ablation experiments of our proposed dual-stream MIL algorithm from three aspects: the role of the dual-stream aggregator, the role of using max-pooling as the top instance, and the role of attention stream. Formally, we compare our proposed aggregator and its counterparts: the model without the aggregator (N-ADDMIL); the top instance of the proposed aggregator obtained by mean pooling (M-ADDMIL); dual-stream aggregator consisting of mean pooling and max pooling (MM-ADDMIL); the model utilizes the attention structure proposed by Ilse *et al.* [30] (A-ADDMIL). We evaluate these five models under the same evaluation system and report the results in Table IV.

As shown in Table IV, our proposed dual-stream MIL aggregator comprehensively improves the performance by comparing with counterparts. By comparing the N-ADDMIL

TABLE IV
EVALUATION OF THE PROPOSED METHOD AND ITS COUNTERPARTS

Method	Accuracy	Precision	Recall	F1
N-ADDMIL	0.702	0.723	0.686	0.714
M-ADDMIL	0.693	0.656	0.636	0.644
MM-ADDMIL	0.706	0.695	0.676	0.684
A-ADDMIL	0.726	0.720	0.700	0.709
ADDMIL	**0.747**	**0.724**	**0.745**	**0.723**

models we can find that adding ADDMIL pooling can obviously and consistently improve the accuracy. In M-ADDMIL, we use the mean-pooling operation to get the top instance of the aggregator. The results show that the performance of M-ADDMIL is the worst among all counterparts, which demonstrates the effectiveness of using max-pooling as the top instance. For a fair comparison, we also consider replacing the branch of the dual-stream aggregator for ablation experiments without changing other conditions. We replace the second stream of our proposed dual-stream aggregator with mean-pooling and the attention mechanism proposed by Ilse *et al.* [30] respectively. These results demonstrate the reliability and superiority of the attention mechanism of the two-stream aggregator.

V. DISCUSSION

The previous state-of-the-art methods are usually supervised learning, and they default the label of each instance (frame or clip of video) to inherit the label of the subject, that is, the label of the bag. Therefore, it is assumed that a high symptom of depression is exhibited in each instance. However, such a processing would introduce too many noisy labels and prevents the model from being trained effectively. The multiple instance learning method we use still achieves fairly good results in this context, providing a new idea for existing depression recognition.

Traditional multiple instance methods based on max-pooling and mean-pooling do not get good results because they only consider instance-level information. The attention-based multiple instance method takes into account the dependencies between the embedding of the instances and the bag, and achieves better results. While our proposed ADDMIL uses non-local operations to additionally consider the dependencies of other instances to the top instance, thus obtaining more complementary contextual information and achieving the state-of-the-art performance.

It is worth noting that our method does not outperform the MSN proposed by Melo *et al.* [7] on all metrics, probably because the 3D convolutional network can extract richer spatiotemporal features. However, our proposed aggregation method can be applied to 2D models with almost no additional parameters and can achieve comparable performance, which means it can be applied to more application scenarios.

The attention weights generated by our algorithm reflect the similarity of all instances in the subject viewing stimulus material to the top instance that is most likely to represent the level of depression of the label. Each of these instances represents the corresponding facial affective state feature. Therefore, the attention weight generated by ADDMIL reflects the correlation between the subject's affective state and the depression level label to a certain extent. The results in Figure 3 suggest that depressed patients may only display affective state associated with depressive level for a small fraction of the time. In the field of emotion recognition using physiological signals, a weakly supervised learning work [58] has also shown that subjects' emotion labels while watching videos are represented only by the most salient or recent emotion. In addition, by analyzing the attention weights between the two groups, we found that depression subjects and healthy subjects have significant differences in their responses to negative video stimulation. This suggests that viewing negatively stimulating videos can serve as a novel experimental paradigm for automatic depression detection.

In view of the widespread spread of the COVID-19 around the world, the prevalence of depression is even more pronounced [59]. However, during the COVID-19 epidemic, clinical interviewing of depressed patients may be limited, which will hinder the treatment and recovery of patients with depression. Online treatment for depression may therefore be more common. Our proposed method demonstrates the ability to identify patient symptoms based on captured video data reflecting the severity of disorder, which can be used an adjunct depression assessment method for patients and provide treatment information. In addition, our model only uses 2D deep network and extra few aggregation operations thus saving a lot of network parameters and GPU memory compared to 3D deep model. Hence, our method is expected to serve most people on smartphones as an APP for depression self-monitoring in the future, which is of great significance for improving the accuracy of initial diagnosis of depression and saving medical resources. We plan to use cloud servers that provide GPU computing to handle model computation and inference, and are prepared to further lighten our models without compromising model performance.

VI. CONCLUSION

In this study, we introduce a novel multiple instance learning framework for depression detection via facial expression videos. Our proposed work is the first attempt at weakly supervised learning on raw video data, which considers facial expressions as a process of change and focuses on useful state information. Our method beats commonly used MIL methods and performs comparable results with the previous state-of-the-art methods. By analyzing the attention weights between video frames, we found that depression patients and healthy subjects had significant differences in their responses to negative video stimulation. We hope that our experiments can also contribute to the detection of other psychiatric disorders through facial expressions in the future work.

REFERENCES

[1] J. L. Sotelo and C. B. Nemeroff, "Depression as a systemic disease," *Personalized Med. Psychiatry*, vols. 1–2, pp. 11–25, Mar. 2017.
[2] S. Evans-Lacko *et al.*, "Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the WHO world mental health (WMH) surveys," *Psychol. Med.*, vol. 48, no. 9, pp. 1560–1571, Jul. 2018.

[3] H. Ellgring, *Non-Verbal Communication in Depression*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[4] D. J. Kupfer, E. Frank, and M. L. Phillips, "Major depressive disorder: New clinical, neurobiological, and treatment perspectives," *Lancet*, vol. 379, no. 9820, pp. 1045–1055, Mar. 2012.

[5] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 578–584, Oct./Dec. 2017.

[6] Z. Jiang, S. Harati, A. Crowell, H. S. Mayberg, S. Nemati, and G. D. Clifford, "Classifying major depressive disorder and response to deep brain stimulation over time by analyzing facial expressions," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 2, pp. 664–672, Feb. 2020.

[7] W. C. deMelo, E. Granger, and A. Hadid, "A deep multiscale spatiotemporal network for assessing depression from facial dynamics," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1581–1592, Jul. 2020.

[8] M. Valstar *et al.*, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2013, pp. 3–10.

[9] M. Valstar *et al.*, "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 3–10.

[10] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[11] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 9, 1996.

[12] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2013, pp. 21–30.

[13] H. Meng and N. Pears, "Descriptive temporal template features for visual motion recognition," *Pattern Recognit. Lett.*, vol. 30, no. 12, pp. 1049–1058, Sep. 2009.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[15] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. 6th ACM Int. Conf. Image video Retr. (CIVR)*, 2007, pp. 401–408.

[16] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: A multimodal approach," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, 2013, pp. 11–20.

[17] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 7, pp. 1432–1441, Jul. 2015.

[18] T. R. Almaev and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 356–361.

[19] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 73–80.

[20] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble CCA for continuous emotion prediction," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge (AVEC)*, 2014, pp. 19–26.

[21] M. Al Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 262–268, Jan. 2021.

[22] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: Generic features for video analysis," 2014, *arXiv:1412.0767*.

[23] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 467–474.

[24] W. C. de Melo, E. Granger, and M. B. Lopez, "MDN: A deep maximization-differentiation network for spatio-temporal depression detection," *IEEE Trans. Affect. Comput.*, early access, Apr. 12, 2021, doi: 10.1109/TAFFC.2021.3072579.

[25] W. C. de Melo, E. Granger, and A. Hadid, "Depression detection based on deep distribution learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4544–4548.

[26] Q. Chen, I. Chaturvedi, S. Ji, and E. Cambria, "Sequential fusion of facial appearance and dynamics for depression recognition," *Pattern Recognit. Lett.*, vol. 150, pp. 115–121, Oct. 2021.

[27] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.

[28] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2016, pp. 2424–2433.

[29] N. Hashimoto *et al.*, "Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3852–3861.

[30] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.

[31] Z. Han *et al.*, "Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2584–2594, Aug. 2020.

[32] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14318–14328.

[33] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, "C-MIL: Continuation multiple instance learning for weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2199–2208.

[34] Z. Luo *et al.*, "Weakly-supervised action localization with expectation-maximization multi-instance learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 729–745.

[35] J. Yao, X. Zhu, and J. Huang, "Deep multi-instance learning for survival prediction from whole slide images," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2019, pp. 496–504.

[36] P. Chikontwe, M. Luna, M. Kang, K. S. Hong, J. H. Ahn, and S. H. Park, "Dual attention multiple instance learning with unsupervised complementary loss for COVID-19 screening," *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102105.

[37] L. Xie, D. Tao, and H. Wei, "Joint structured sparsity regularized multiview dimension reduction for video-based facial expression recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 2, pp. 1–21, Jan. 2016.

[38] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2014.

[39] C. Wu, S. Wang, and Q. Ji, "Multi-instance hidden Markov model for facial expression recognition," in *Proc. IEEE Int. Conf. Workshops Automat. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–6.

[40] Y. Fang and L. Chang, "Multi-instance feature learning based on sparse representation for facial expression recognition," in *Proc. Int. Conf. Multimedia Modeling*. Cham, Switzerland: Springer, 2015, pp. 224–233.

[41] L. Xie, D. Tao, and H. Wei, "Early expression detection via online multi-instance learning with nonlinear extension," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1486–1496, May 2018.

[42] J. Wu, B. Yang, Y. Wang, and G. Hattori, "Advanced multi-instance learning method with multi-features engineering and conservative optimization for engagement intensity prediction," in *Proc. Int. Conf. Multimodal Interact.*, 2020, pp. 777–783.

[43] A. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "EmotiW 2020: Driver gaze, group emotion, Student engagement and physiological signal based challenges," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 784–789.

[44] A. Salekin, J. W. Eberle, J. J. Glenn, B. A. Teachman, and J. A. Stankovic, "A weakly supervised learning framework for detecting social anxiety and depression," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 2, pp. 1–26, Jul. 2018.

[45] Z. Ren, J. Han, N. Cummins, Q. Kong, M. D. Plumbley, and B. W. Schuller, "Multi-instance learning for bipolar disorder diagnosis using weakly labelled speech data," in *Proc. 9th Int. Conf. Digit. Public Health*, 2019, pp. 79–83.

[46] Y. Wang *et al.*, "Automatic depression detection via facial expressions using multiple instance learning," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1933–1936.

[47] D. Rymarczyk, A. Borowa, J. Tabor, and B. Zielinski, "Kernel self-attention for weakly-supervised image classification using deep multiple instance learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2021, pp. 1721–1730.

[48] J. Feng and Z.-H. Zhou, "Deep MIML network," *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, Feb. 2017, pp. 1–7.

[49] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1713–1721.

[50] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognit.*, vol. 74, pp. 15–24, Feb. 2018.

[51] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[52] D. E. King, "Dlib-Ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jan. 2009.

[53] D. H. Kim and B. C. Song, "Contrastive adversarial learning for personindependent facial emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 7, 2021, pp. 5948–5956.

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.

[55] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan./Mar. 2017.

[56] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[57] L. He, J. C.-W. Chan, and Z. Wang, "Automatic depression recognition using CNN with attention mechanism from videos," *Neurocomputing*, vol. 422, pp. 165–175, Jan. 2021.

[58] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, "Weakly-supervised learning for fine-grained emotion recognition using physiological signals," *IEEE Trans. Affect. Comput.*, early access, Mar. 10, 2022, doi: 10.1109/TAFFC.2022.3158234.

[59] C. K. Ettman, S. M. Abdalla, G. H. Cohen, L. Sampson, P. M. Vivier, and S. Galea, "Prevalence of depression symptoms in U.S. adults before and during the COVID-19 pandemic," *JAMA Netw. Open*, vol. 3, no. 9, Sep. 2020, Art. no. e2019686.