Recognition of Depression from Video Frames by using Convolutional Neural Networks

Jianwen WANG, Xiao SHA*
Department of Computer Science, Hebei University of Water Resources and Electric Engineering, Hebei 061001, China

***Abstract*—The disturbances of the mood are relevant to the emotions. Specifically, the behaviour of persons with disturbances of mood, like the depression of the unipolar, displays a powerful correlation of the temporal by the emotional girths of the arousal and the valence. Moreover, the psychiatrists and the psychologists take into account the audible signs of the facial and the audible signs of the voice when they assess the condition of the patient. Depression makes audible behaviours like weak expressions, the validation of the contact of the eye and the use of little flat-voiced sentences. Artificial intelligence has combined various automated frameworks for the detection of depression severity by using hand-crafted features. The method of deep learning has been successfully applied to detect depression. In the current article, a federate architecture, which is the network of the neural of the deep convolutional basis on the attention of global, is proposed to diagnose the depression. This method uses CNN with the attention mechanism and also uses the integration of the weighted spatial pyramid pooling for the learning of the deep global representation. In this method, two branches are introduced: the CNN based on local attention focuses on the patches of the local, while the CNN based on global attention attains the universal patterns from the whole face area. For taking the data of the supplementary among two parts, a CNN basis on the local-global attention is proposed. The designed experiments have been done in two datasets, which are AVEC2014 and AVEC2013. The results show that our presented approach can extract the depression patterns from the video frames. Also, the outcomes display that our presented approach is superior to the best methods based on the video for the detection of depression.**

***Keywords*—Deep learning; depression recognition; Convolutional Neural Network (CNN); attention mechanism**

I. Introduction

By 2020, depression had the 4-th rank among the most earnest issues of the health of the mentally [1]. Generally, it does temperate damage to the life of the individual. Also, it has a special effect on the society and the family. In several instances, depression may cause self-annihilation. Therefore, it is essential to discover an impressive solution for the diagnosis of depression and the treatment of depression of the clinical.

In recent years, a multitude of approaches have been proposed based on the different perspectives to help psychologists or doctors with the detection and treatment of clinical depression; these methods have mainly used emotional computations, machine learning communities, computer vision, etc. for estimating the depression's severity on the basis of the audiovisual cues, the common methods usually include three sequential methods: a) The extraction of the feature, b) The aggregation of the feature, and c) The regression (the classification). The extraction of the feature acts as an important task in the detection of depression from the videos. It is very important to extract a distinctive feature descriptor for the diagnosis and the estimation of depression [1]. Due to the extraction of the feature, the approaches can be hastily distributed to the features of the hand-crafted and the features of the deep learning.

The features of the hand-crafted use the knowledge domain for the designing of the features which are relevant as closely to the depressive signs [2], [3]. Although the hand-crafted features representation is considered for the obtention of the superior performance for the depression severity assessment, the below subjects have been related by the researchers. First, the exploitation of the features of the hand-crafted is time-consuming because these features require particular knowledge. The patterns of the binary of the local consist of 3 planes of the orthogonal [4], and they are heavy in terms of computational. Second, the hand-crafted features have been criticized due to the lack of related significant information on the patterns of depression [5].

Newly, the learned deep features with the use of CNNs have been applied widely to represent the deep features, and they have done great in depression diagnosis [6]. DepressNet [6] is a new framework for the learning of annotated depression representations. The selected methods of CNN (e.g., GoogleNet [7], AlexNet [8], VGG-Net [9], etc.) are pre-trained in the big datasets of the picture of the facial [10] and then it is trained in the dataset of AVEC2013 [11] and the dataset of AVEC2014 [12] with the use of fine-tuning. Its performance surpasses the many approaches to the diagnosis of depression based on the video. [13] uses the composition from the RNN [14] and the 3D-CNN [15] for the learning of the representation of the consecutive features of the spatial-temporal in 2 various scales of the areas of the face. [6], [13] adopt the model of the deep of the pre-trained for the fine-tuning of 2 datasets of the depression for the estimation of the depression.

In general, the diagnosis of depression is a problem of regression or a problem of classification according to machine learning. The purpose of the dataset of AVEC2013 and the dataset of AVEC2014 is prediction of the depression scores. It is proposed that almost all of the non-literal behaviours in the interaction of humans are the anent of the area of the face [16]. Regarding the estimation of depression based on the video, the salient area of the face is applicable for anticipating depression severity, as has been proposed in [6].

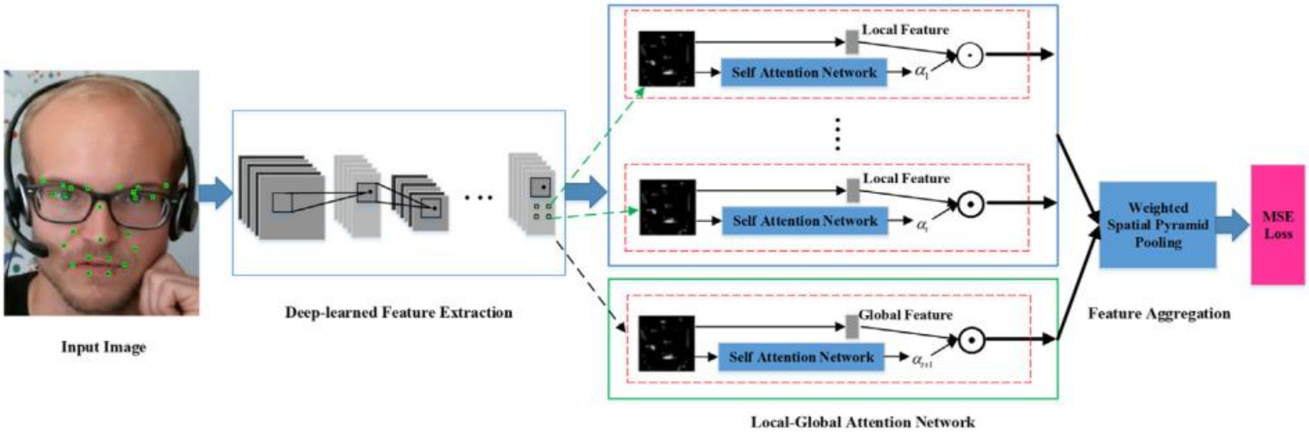

Fig. 1. Framework of our presented method to detect depression.

In the current article, the exploration of the techs on the basis of the facial look for depression diagnosis is concentrated. In order to deal with the mentioned problems, a new approach is proposed for depression detection by using the face video frames, which is called the DCNN, based on the attention of local-global. As displayed in Fig. 1, our proposed method includes 3 parts: i) The module of the extraction of the deep features learned from Depressed-CNN, ii) The convolutional neural network based on the local-global attention, iii) The module of the weighted spatial pyramid pooling (i.e., WSPP). The purpose of the universal attributes is to explain the set of depression-specific patterns, when the local attributes focus on the record of the specific patterns in the patch regions, which can extract the distinctive features in the patches of the prominent maps of the feature. The information extraction of the features of the local and the features of the global is crucial for the better performance of the depression diagnosis.

In order to clearly state the differences between the proposed method and the previous methods, we summarize the key contributions of the present paper as follows: 1) An end-to-end framework with the deep global-local attention is proposed, which effectively uses the face dynamics as a non-verbal metric to estimate the severity of the depression scale, which it has been neglected in the previous methods. 2) To encode the robust feature representations, a sophisticated CNN-based feature extraction network is designed. This network preserves the valuable and the distinctive features useful for the analysis of depression. 3) A CNN with a self-attention network effectively describes the discriminative patterns of the faces. By adopting the attention mechanism, the global-local attention-based CNN can automatically preserve the valuable feature and filter out the redundant face information. The continuation of the current article is as follows: Section II provides an overview of previous works. In Section III, our proposed approach is provided. In Section IV, the used datasets and the experiment results are provided. Section V discusses the general conclusions and future perspectives.

II. An Overview of Related Works

Many works have been done for the analysis of depression on the dataset of AVEC2013 and the dataset of AVEC2014. In the following, we introduce some presented approaches for the analysis of depression in the video frames. Liu et al. [17] have designed a region-based global network with partial attention and relational attention that this network learns the relationship between the partial features and the features of the global. In [18], the authors have introduced a framework with the use of CNN and the mechanism of attention to automatically detect depression by facial changes, whose performance overtakes most methods of facial depression detection. By focusing on the attention mechanisms and by paying attention to the facial details, this method has achieved promising results. In [19], authors have presented the deep network of the regression for the learning of the representation of the depression features visually and interpretatively, and its results show that the area near the eye plays a significant task in the diagnosis of depression. Al Jazaery and Guo [20], with the use of the 3D-CNN and the RNN, have learned as automatically the features of the spatial-temporal areas of the face in 2 various scales that can model the information of the local of the spatial-temporal and the information of the global of the spatial-temporal by the steady expressions of the facial to forecast the depression level.

On [2], the local features of the phase of the quantization by 3 planes of the orthogonal are extracted by using the coding of the sparse, and then, they are displayed with the discriminant map and by the decision surface fusion method for the generation of the features of the top-level. The regression of the vector of the support is adopted for evaluation of the severity of the depression. On [21], a new feature of the temporal dynamic, which is called the strong patterns of the binary of the average of the local by 3 planes of the orthogonal, is applied to provide the expressions of the dynamic of the temporal of the face. For the creation of a representation of the vector of the feature, the vector of Fisher of the Dirichlet process learns a richer intermediate representation from the MRLBP-TOP features in the subsequences. Next, for every sample of the video, a representation of the discriminative is generated by using the statistical aggregation approaches. In [22], the authors train an architecture of CNN for a combination of the appearance of the facial and the dynamics of the facial to evaluate the depression diagnosis scale. The authors have related the superior outcomes over the other approaches based on the visual.

In [23], the authors design a system of artificial intelligence to estimate the depression scale. This system can combine the pattern of the supplementary among the features of the hand-crafted and the features of the deep learning. In the cues of the visual, the features of deep learning are exploited, and these features contain some related discriminative information to depression. In the cues of the audio, the features exploit the descriptors of the spectral with low level and the coefficients of the brain of the frequency of the Mel to take the expression of the vocal by the clips of the audio. The temporal movement in the space of the various features is defined with the histogram of the history of the dynamic from the feature. [6] propos Depress-Net, which is the deep network of the regression for prediction of the severity of the depression by the alone images. The map of the activation of the depression is applied in order to show the areas of the salient from the image of the face to determine the depression scale. In the meantime, the authors have designed a Depress-Net of the multi-area for modelling the various patterns of the various areas to better the total outcomes. Vast tests have been done on the dataset of AVEC2013 and the dataset of AVEC2014. Its efficiency has shown that the presented method outperforms the best visual-based methods of depression detection.

In [24], the authors extract the features of the global-local of the 3D-CNN to enhance the method's efficiency. The presented model is equipped with the 3D pooling of the average of the global for the representation of the patterns of the temporal-spatial for the diagnosis of depression. The empirical outcomes display that the integration of the features of the local and the features of the global of the 3D-CNN achieves promising efficiency. On [25], the authors have proposed to exploit automatically the basic human behaviours as the descriptors with low-dimensional from every frame. Two representations of the feature of the spectral, namely the spectral heat-maps and the vectors of the spectral, have been presented for the capturing of the associated multiscale patterns with the depression. The authors have presented these two spectral representations for the prediction of depression by using CNN and artificial neural networks. In [26], the 2-stream framework of the spatial-temporal for the depression diagnosis is presented. Eke, the researchers present the time-averaged integration method for the generation of the time-slice features. The tests on the dataset of AVEC2013 and the dataset of AVEC2014 have shown that their presented framework achieves comparable performance for depression detection.

III. Our Presented Approach

In the current part, the details of our presented approach are described. The framework of the diagnostic of depression, which is the deep and end-to-end, is shown in Fig. 1. First, the area of the face is cropped by the clips of the video with the use of the OpenFace toolbox [27]. Then, the usual CNN for the extraction of the feature is implemented, and this CNN obtains the maps of the feature of the face. In order to filter the features of the additional, the various networks of self-attention are introduced in the maps of the feature of the local and the maps of the feature of the global. WSPP is adopted to create the scale-variable features representation on the multiscale feature maps. Finally, two layers of the fully connected and a layer of the loss of the MSE are used to predict the severity of the depression. On each of the below sub-parts, the details of every part from our presented approach are provided.

A. *The Extraction of Feature for the Obtention of the Face Feature Maps*

The CNNs have been applied widely, and it has been proven that these networks are impressive in extracting the features in the field of emotional computations, like the recognition of the expression of the depression diagnosis and so on. To overcome the small datasets, our presented method for the analysis of depression is inspired by the described method in [28] with the un-deep frameworks. Since the deep frameworks, we are able to model a distinct presentation to predict the depression severity range. To extract the deep learning features, the presented shape of the framework of Depressed-CNN is shown in Table I. Meantime, the Depressed-CNN architecture is shown in Fig. 2.

TABLE I. Our Proposed Configuration for the Depressed-CNN Architecture

Name	Input	Operation	Kernel	Output
Conv1	$224 \times 224 \times 3$	Convolution	3×3, ReLu	$224 \times 224 \times 64$
Pool1	$224 \times 224 \times 64$	Pooling	2×2	$112 \times 112 \times 64$
Conv2	$112 \times 112 \times 64$	Convolution	3×3, ReLu	$112 \times 112 \times 128$
Pool2	$112 \times 112 \times 128$	Pooling	2×2	$56 \times 56 \times 128$
Conv3	$56 \times 56 \times 126$	Convolution	3×3, ReLu	$56 \times 56 \times 256$
Pool3	$56 \times 56 \times 256$	Pooling	2×2	$28 \times 28 \times 256$
Conv4_1	$28 \times 28 \times 256$	Convolution	3×3, ReLu	$28 \times 28 \times 512$
Conv4_2	$28 \times 28 \times 512$	Convolution	3×3, ReLu	$28 \times 28 \times 512$
CMP	$28 \times 28 \times 512$	Pooling	4×4	$28 \times 28 \times 128$
Con	CMP+Conv4_2	Concatenate	/	$28 \times 28 \times 640$

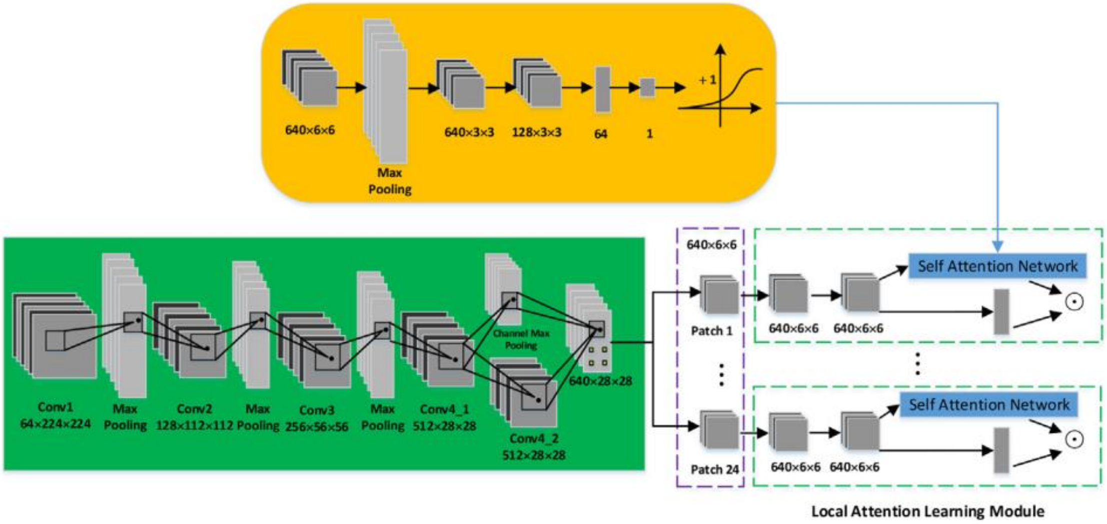

Fig. 2. Our detailed view of CNN based on the local attention.

The features of deep learning are exploited by Depressed-CNN as below. The image size input of the face is equal to 224×224 by three channels of colour. Inspired by [29], for total convolution layers, a filter by a little kernel of 3×3 is used to encode the left/right concept and the up/down concept for the extraction of the information of the spatial-temporal. After 3 operations of the convolutional and after three max-pooling operations, $Conv4\_1$ will be $[512 \times 28 \times 28]$. In the proposed method, the max-pooling is done on a window of 2×2 by $stride = 2$. To prevent the loss of information and also to preserve the distinct feature, the max-pooling of the channel (CMP) is used for the pooling of the map of the feature on the direction of the channel by the size of the kernel equal to $4, stride = 4, pad = 0$. The output is a 3D map of the feature with a size equal to $[128 \times 28 \times 28]$. This structure is similar to this form because the popular max-pooling calculates the value of the maximum on the direction of the spatial while the max-pooling of the channel calculates the value of the maximum on the direction of the channel. In addition, for the creation of a strong representation of the feature, the map of the feature of $Conv4\_2$ is created on $Conv4\_1$. Next, CMP and the map of the feature of $Conv4\_2$ are concatenated together to achieve the feature of the final by size equal to $[640 \times 28 \times 28]$.

B. Convolutional Neural Network based on Local Attention

The various components of the image of the face have a great contribution to the diagnosis of depression. Inspired by [27], the area of the face is cropped to the various patches for the capturing of the representation of the distinct features for depression analysis. The CNN based on the local attention consists of two main steps: the generation of the patch and the capture of the salient feature. In the below, these 2 stages are described in the detail.

-The Generation of the Patch: For the analysis of the depression by using the area of the face, the patches of the local may have discriminating features to estimate the depression severity. As the popular scheme of the extraction of the feature of deep learning, the CNN restriction is the learning of the geometric transformations. Since some facial muscles contain specific information for the state recognition that is relevant as closely to the depression [30], the area of the face is cropped to the distinct patches. Also, in [31], [32], the researchers believe that the information of the multi-view from the areas of the salient is significant for image recovery. For the mentioned purpose, it is suggested that the proposed method adopts the various patches to take the distinct representations to detect depression. In the current article, the toolbox of OpenFace is used for the detection of the face region and also for the aligning of each frame from the video sequences with the size equal to 224×224 in three colour channels. Next, 68 landmarks of the face are recognized (Fig. 3). In order to find the effective patches of the face which are relevant as closely to the depression, 16 points are selected from 68 points, and these points cover the distinct face patches. Then, 8 points, which are covered on the eyes and the cheeks of the face, are recalculated. In total, 24 face patches were extracted. Our presented approach is displayed in Fig. 3.

The stages of the process of processing are as follows: (a) The OpenFace toolbox is used to discover the face zone and fix the face on every image from the trail of the video. 26 points are selected. These points cover the main areas of the face, namely, the eyes, the nose and the mouth. The elected points have the following indexes: $18, 19, 20, 37, 38, 39, 41, 42, 22, 23, 25, 26, 27, 44, 45, 46, 48, 47, 28, 30, 49, 51, 59,$ 53, 55, 57. (b) 4 pairs of the face landmarks (38.20), (41.42), (45.25), (48.47) are taken. Finally, 16 points from 26 points are recalculated. (c) The middle point of each pair of the points is calculated. (d) For the detection of the area around the mouth, two pairs of points (59.18) and (57.27) are selected, and then the middle points are calculated. The indices of the middle point are $22, 21$. Next, 2 points, which have a similar space from the corners of the mouth, are calculated. For the coordinates of the goal edges of the direction of left, we describe it as $(U.V) = (U_{left} - 16.V_{left} - 16)$. The coordinates of the point of the target in the right direction it is described as $(U.V) = (U_{right} - 16.V_{right} - 16)$. The interval between the corner of the mouth and the edge of the target is calculated, and then the indices are determined as $24,23$. (e)

Next, it selects 24 landmarks of the face, which cover the key areas of the face. (f) Due to the location of 24 points, 24 patches are cut. The patches are created by the image of the face. Nevertheless, in the proposed method, the generation of the patch is done in the maps of the feature for obtention of the areas of the salient.

The Registration of the Features of the Salient: As displayed in the box of green in Fig. 2, the process of the generation of the patch is performed with the use of the maps of the feature of CNN instead of the images of the face. This point is to maximize the utilization of the operations of the convolution and the strengthening of the fields of the receptive of the neurons, which reduces the model size. The output of the method of the extraction of the feature is the maps of the feature with a size equal to $[640 \times 28 \times 28]$. For the patch generation, 24 local zones equal to $[640 \times 6 \times 6]$ are obtained. To achieve the attributes of the salient, we apply the module of the learning of the attention of the local by CNN based on the local attention to learn the features of the patches of the face automatically. The module of learning the attention of the local is displayed on the box of yellow on the midmost from Fig. 2 by 2 dashed rectangles of green. On every module of the learning of the local attention of the patch-specific, after the generation of the patch, the created maps of the feature are entered in 2 layers of the convolution. Next, the second maps of the feature are entered into 2 branches. The 1-th branch considers the maps of the feature as the features of the local level of the vector. For the 2-th branch, a network of self-attention is applied for the focus on the distinct representation areas by the patches of the spatial. Next, the feature of the patch is recomputed by the vector of the weight.

Formally, let P_{a_j} is j-th patch of the map of the feature by the size equal to $[640 \times 6 \times 6]$. $\hat{P}_{a_j^1} = f(P_{a_j})$ is the 1-th map of the feature in the dashed rectangle of green on the top of Fig. 5, which map has a size equal to $[640 \times 6 \times 6]$. After a filter with a size equal to 1×1, the 2-th map of the feature is $\hat{P}_{a_j^2} = f(P_{a_j})$, which this map has a size equal to $[640 \times 6 \times 6]$. f represents the operation of the convolution in the architecture of CNN. Next, the 2-th map of the feature is entered into 2 branches. The 1-th branch converts the 2-th map of the feature to a vector of the local feature. Suppose φ_j is the map of the feature of the local that takes the 2-th map of the feature as the input. φ_j can be defined as the follows:

$$\varphi_j = \varphi(\hat{P}_{a_j^2}) \tag{1}$$

φ is the operation of the vector transformation. The 2-th branch is the network of self-attention. This network consists of an operation of the max-pooling, an operation of the convolution, 2 layers of the fully connected and an operation of the sigmoid. The function of the sigmoid is applied to limit the output scope α_j in 0-1. In it, 0 represents a related patch, and 1 displays a critical patch for the detection of depression. The weight α_j can be described as the follows:

$$\alpha_j = \omega_j(\hat{P}_{a_j^2}) \tag{2}$$

Where α_j is the scalar, and ω_j displays the operation of the network of the self-attention. After the operations of 2 branches, α_j is applied in the feature of the local φ_j to create a distinct feature:

$$\rho_j = \alpha_j \cdot \varphi_j \tag{3}$$

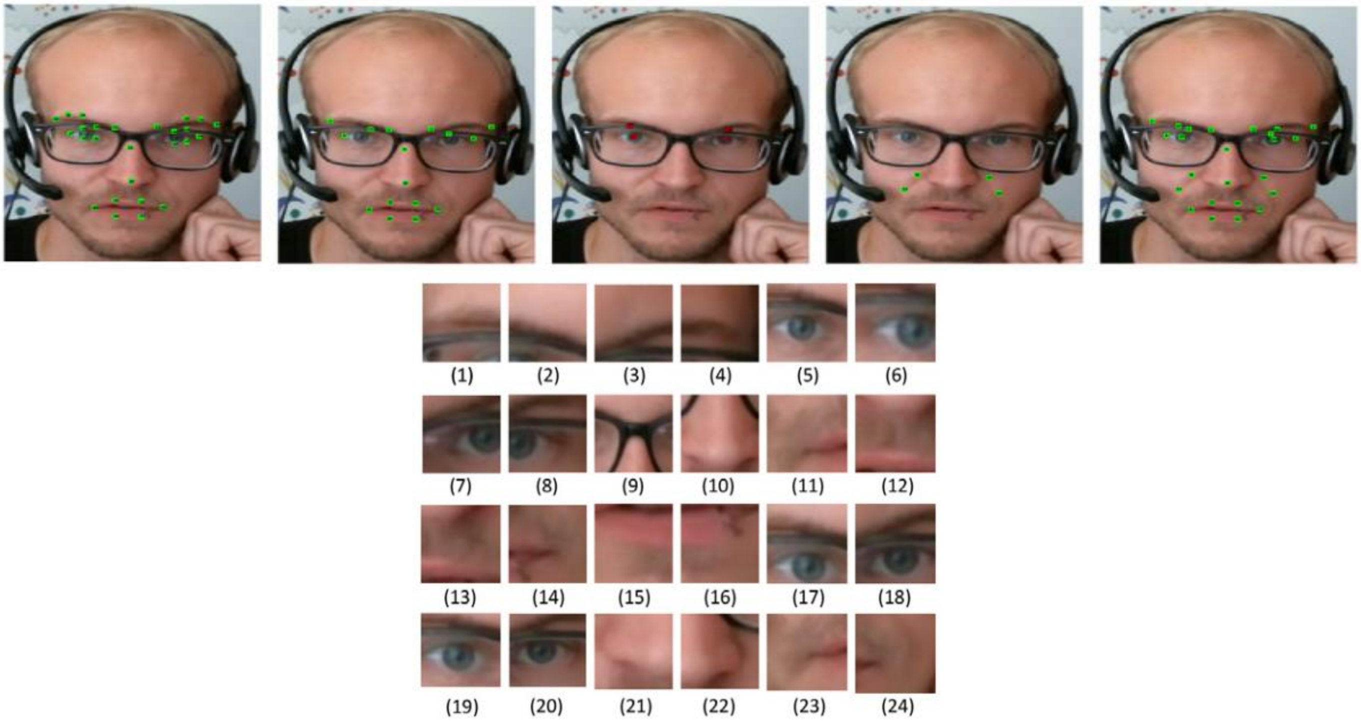

Fig. 3. An example of the face patch generation by using the mentioned patch generation process.

The output feature contains a distinct representation of the depression. Specifically, each module of local attention learning is weighted with weights, which are learned automatically with the network of self-attention. For CNN based on local attention, it is a framework of the depression of the local end-to-end with the below parts: the extraction of the deep learned features from CNN, the generation of the patch and the mechanism of the attention. The feature basis on the patch, which is learned by the proposed deep network, can discover the visual sinking pattern in the face region, and with it, the automatic estimation of the depression is feasible.

C. Convolutional Neural Network based on Global-Local Attention

In this section, we first provide a description from CNN based on global attention, and then, we state the final proposed method, namely CNN based on global-local attention. As described above, CNN, based on local attention, can automatically learn the distinctive features by using the mechanism of attention to analyze depression. Nevertheless, patches of the CNN based on the local attention may miss some additional information, which this information includes the face images and the general information of the semantics for the pattern of the depression. Therefore, for improvement of performance and also for the learning of the deep information of the semantics, the CNN based on local-global attention is proposed. For the part of the extraction of the feature, a similar operation of CNN based on local attention is used. In the presented implementation, it is proposed that the module of the learning of the attention of the global be used to represent the information of the semantics of the global in the depression diagnosis (the dashed rectangle of red in Fig. 4.

The module of the learning of global attention includes an operation of the max-pooling, an operation of the convolution and 2 operations of the branch. Ere entering to 2 branches, the map of the feature $Conv6$ can be defined as g by a size equal to $[512 \times 14 \times 14]$. The 1-th branch converts the maps of the feature g to the vectors of the feature of the global. Suppose Ψ is a feature of the global, and it is defined as the follows:

$$\Psi_{j+1} = \varphi(g) \tag{4}$$

φ is the operation of the vector transformation. The 2-th branch is the network of self-attention that includes an operation of the max-pooling, an operation of the convolution, 2 layers of the fully connected and an operation of the sigmoid. The weight α_{j+1} can be described as the follows:

$$\alpha_{j+1} = \omega_{j+1}(g) \tag{5}$$

Where α_{j+1} is the scalar, and ω_{j+1} represents the operation of the network of the self-attention. α_{j+1} is weighted in the feature of the global φ_{j+1} to obtain the feature containing the useful information ρ_{j+1}:

$$\rho_{j+1} = \alpha_{j+1} \cdot \varphi_{j+1} \tag{6}$$

In addition, to encode the complementary representations in CNN based on the local attention and in CNN based on the global attention, a final end-to-end architecture, which is called the CNN based on the local-global attention (the rectangle of red in Fig. 5) is proposed to connect this CNNs together.

D. Weighted Spatial Pyramid Pooling

The key purpose of the current article is the evaluation of depression severity. To increase the depression severity indices, the deep representation should capture the distinctive facial features at the various measures. Therefore, an important number of the samples, in total cases of the natural, are required in the training phase. The sources of the spatial of the natural changes of the face include the positions, the facial area size and the angles. In the proposed method, the features of the deeply learned local and the features of the deeply learned global by the mechanism of attention are used to obtain the information of the discriminative directly to diagnose depression. Nevertheless, the movement of the head may make changes in face size within a trail of the image. Thus, the changes may result in the blurring of the face, and this blurring affects the face image clarity.

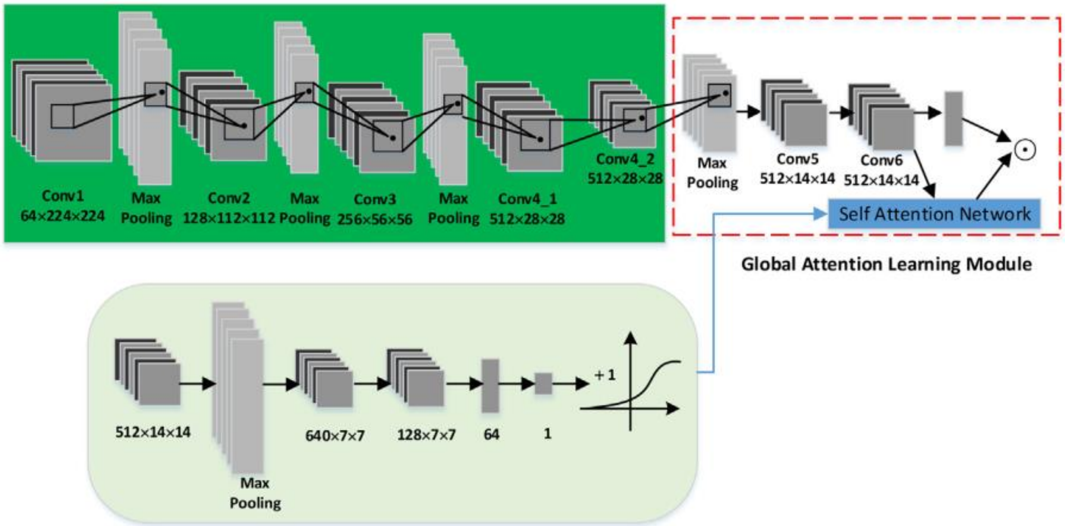

Fig. 4. Our detailed view of CNN based on the global attention.

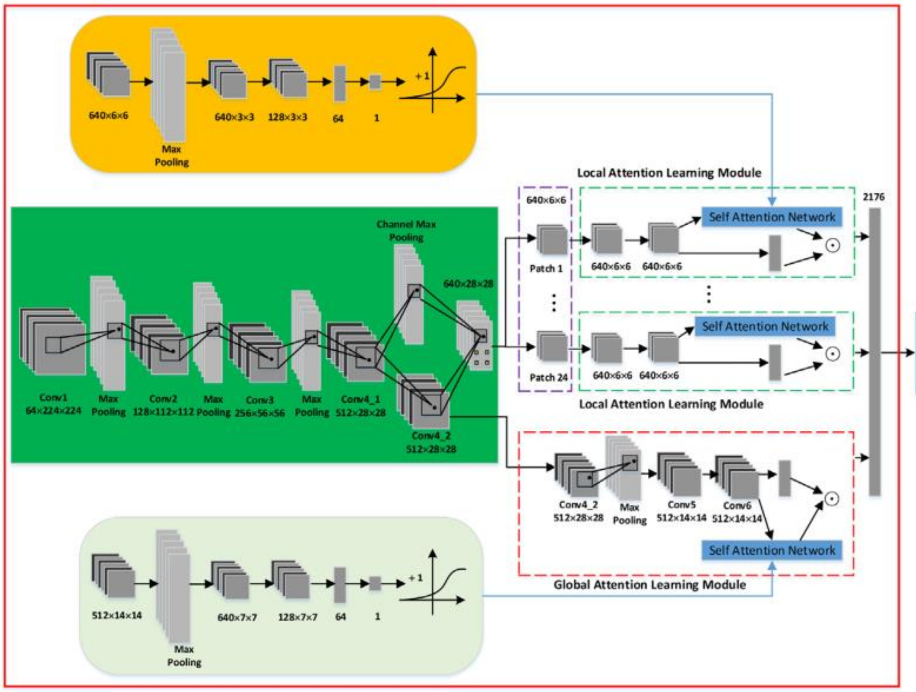

Fig. 5. Our view of the method of CNN based on local-global attention with WSPP.

To achieve a representation of the static scale, a layer of WSPP is applied for the representation of the multiple scales in the output upside of CNN based on local-global attention. The WSPP idea is to segment the map of the feature to the various parts from the scales of the finer-to-coarser, which is finished by the aggregation of the features of the local. The layer of WSPP can better the non-scaling. Also, it reduces the problem of over-fitting. In this paper, we use the definition of the initial weight on the spatial pyramid kernel, which this definition is presented in [33]. The features at the resolutions of the finer are related to the weight of the heavier ones, and the attributes of the coarser solutions are supported by the load of the lower ones. Fig. 6 describes the step of the representation of the feature of WSPP. The output shape of CNN based on the local-global attention is equal to $2048 \times 1 \times 1$.

On each spatial pyramid, the max-pooling is used for the combination of responses of every filter. The WSPP output is equal to the sum of the overall number from the pyramids. The output of CNN based on the local-global attention has the shape of $[batch\_size. 2048]$. A vector of the feature of the final by the size equal to $150D$ is obtained. The size of the window of WSPP is equal to $25 \times\ 151 \times\ 1.102\ \times\ 1.204 \times 1$, and their corresponding stride is equal to $25\ \times\ 151\ \times\ 1.102\ \times\ 1.204 \times 1$. The vectors with the fixed dimensions are fed into the layer of the fully connected.

For a network of the deep, the function of the loss plays an important task in the regression of the target. The analysis of the depression can be considered as a problem of the regression. Thus, in the proposed method, the loss of Euclidean is applied as the function of the loss, and this function is appropriate for our proposed method. The function of the loss of Euclidean L computes the squares sum of the disagreement among the values of the actual and the estimated values. It can be expressed as follows:

$$\mathrm{L} = \frac{1}{2M}\sum_{i=1}^{M} ||\hat{p}_i - p_i||^2 \qquad (7)$$

M is the sample number, and $\hat{p}_i$ displays the architecture output. Also, p_i displays the label. In this way, the final proposed architecture for the depression scale estimation is obtained, and all parts of it are shown in Fig. 5.

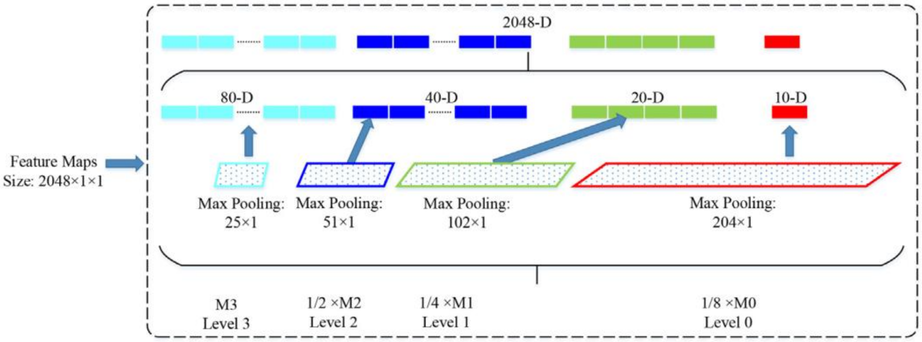

Fig. 6. Our view from WSPP.

IV. The Experiments and the Results Evaluation

In the current part, the details of our used datasets, the performed tests and also, and the obtained outcomes are presented. The Python programming language has been used for the implementation of these experiments. The presented method is implemented on a computer with 8G RAM and Core (TM) i7 CPU 3.0 GHz Intel(R). The network of the convolutional is designed on GPU, and the card used for the graphics on our approach is GEFORCE 840M from NVIDIA.

A. The Used Datasets

To prove the performance of the proposed approach for depression detection, the tests on two datasets are conducted: AVEC 2014 and AVEC 2013. The distribution of the scores of BDI-II on the dataset of AVEC 2013 and the dataset of AVEC 2014 is displayed in Fig. 7. In the dataset of AVEC 2013, 150 clips of the video exist. These clips are taken from 82 participants on the interaction of the computer-human by a webcam and a microphone for the record of the data. The scope of the age for the total people on the dataset is equal to 18 years to 63 years by a mean age equal to 31.5 years and also with a standard deviation equal to 12.3 years. The recorded clips are adjusted to 30 frames every second by a resolution equal to 640×480. The dataset of AVEC 2013 is distributed to 3 partitions: the development, the test and the training. In each partition, this dataset has 50 videos. Every video has a corresponding label with its level of depression intensity, and this level is evaluated based on the questionnaire of BDI-II.

The dataset of AVEC 2014 is the subset of the dataset of AVEC 2013. In it, there are 2 works: Northwind and FreeForm. These two works have 150 clips. In the work of FreeForm, the persons answered multiple questions, like the description of a sorrowful memory in childhood or the expression of popular food. In the work of Northwind, the persons had to study a selective from a fairy tale aurally. Similar to the AVEC 2013 dataset, the AVEC2014 dataset has 3 partitions: the development, the test and the training. The tests are done by using the partition of the training and the partition of the development from two tasks as data of training, and then, the partition of the test is applied for measurement of the model performance.

B. The Experiment Settings and the Evaluation Criteria

To obtain fast convergence and the optimization of the model, we apply the AdamW [34] by the adaptive learning rate strategy. The size of the batch is adjusted to 64, the rate of the dropout is adjusted to 0.2, and the learning coefficient is adjusted to 0.1. Regarding the rate of learning, the tests are done to check the efficiency of our presented approach at different rates.

The efficiency of base approaches in the AVEC 2013 dataset and the AVEC 2014 dataset is evaluated based on two evaluation criteria: the MAE and the RSME. MAE and RMSE are considered as criteria during the experiment for a fair comparison. These criteria are defined as follows:

$$\mathrm{MAE} = \frac{1}{M} \sum_{j=1}^{M} |\hat{p}_j - p_j| \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{M} \sum_{j=1}^{M} (\hat{p}_j - p_j)^2} \tag{9}$$

Where M displays the overall number of samples of the video and p_j and $\hat{p}_j$ represent the actual score and the estimated score of BDI-II from j-th video.

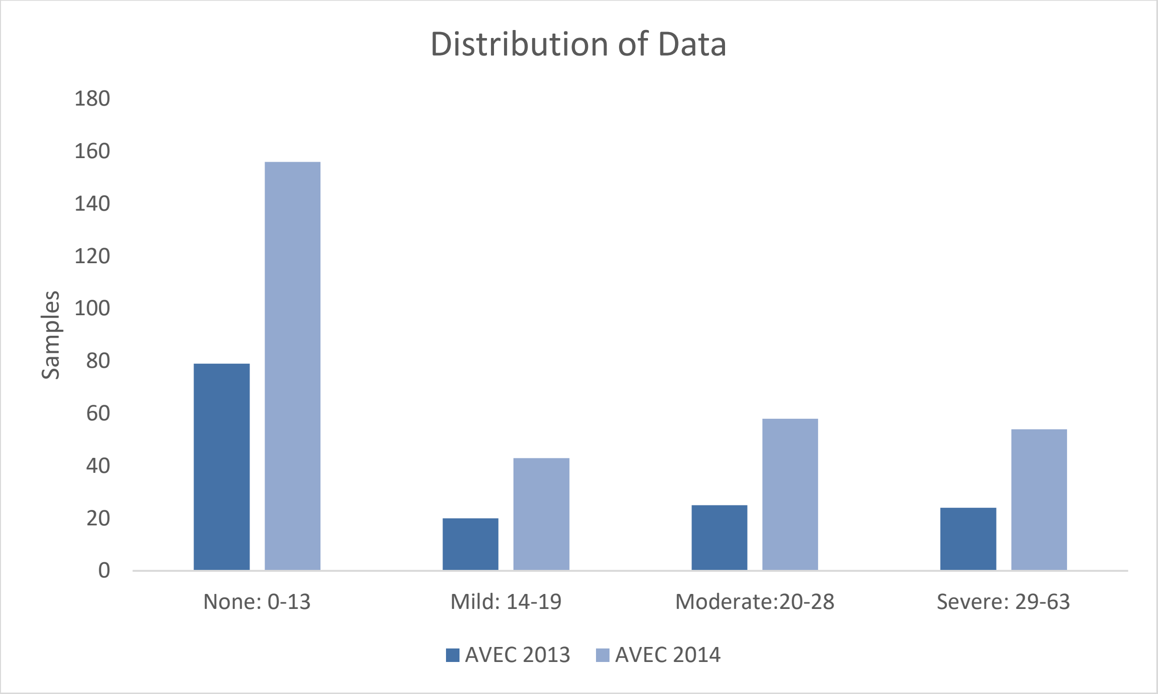

Fig. 7. Distribution of the scores of BDI-II on the dataset of AVEC 2013 and the dataset of AVEC 2014.

C. Evaluation of the Obtained Results

In the current part, first, we conduct an erosion check to investigate the performance of the components of the individual on our proposed approach. Next, our proposed framework is compared by the multiple other methods in this field to demonstrate its promising performance. The initialization of the model by LA-CNN (CMP-) and LA-CNN (CMP+), respectively, indicates the use and the non-use of the CMP technology in CNN based on local attention. The initialization of the model by GA-CNN shows that just the mechanism of the attention of the global is applied to the diagnosis of depression. The residual of the initialization of the model in Table II and Table III combine 2 or 3 separate components to take the information of the supplementary among them. Table II shows the detection results in the partition of the test from the dataset of AVEC2013. It is seen that $G1$ achieves the foremost efficiency. For the dataset of AVEC2014, we conducted various experiments for the verification of the efficiency of our presented approach. In Table III, it can be seen that similar perceptions with the AVEC2013 dataset are obtained. Additionally, from the 2 tables, it can be seen that the various approaches of CNN based on local attention have better performance than the different models of CNN based on global attention and $C1$. The performance in two depression datasets displays which capability of our presented approach is suitable for the evaluation of the depression severity scale from the video sequences. These observations show that with the combination of the components of an individual, the total efficiency is improved more over the use of a component of an individual. This point implies that we need to integrate the models of local and the models of global to diagnose depression. For a fair comparison with the other methods, just $G1$ and $G2$ are adopted to evaluate the depression diagnosis models, which are the foremost outcomes of our presented approach.

Regarding the learning rate, we changed this rate in the scope of 0.000001-0.1 and other parameters are static. As displayed in Tables IV and V, the increment of the rate of learning decreases the efficiency. Since the larger rate of learning leads to poor performance, the minimum rate of learning is appropriate for our presented approach to learn the significant patterns which are relevant to depression closely.

TABLE II. THE EFFICIENCY OF THE VARIOUS COMBINATIONS OF OUR PRESENTED METHOD IN THE AVEC2013 DATASET

Model Setting	RMSE	MAE
LA-CNN (CMP-)	8.74	7.19
LA-CNN (CMP+)	8.65	7.12
A1: LA-CNN (CMP-)+WSPP	8.56	6.81
B1: LA-CNN (CMP-)+WSPP	8.40	6.52
GA-CNN	9.05	7.43
C1: GA-CNN+WSPP	8.98	7.36
D1: LA-CNN(CMP-)+GA-CNN	8.71	7.15
E1: LA-CNN(CMP+)+GA-CNN	8.63	7.02
F1: LA-CNN(CMP-)+GA-CNN+WSPP	8.52	6.80
G1: LA-CNN(CMP+)+GA-CNN+WSPP	8.30	6.48

TABLE III. THE EFFICIENCY OF THE VARIOUS MIXTURES OF OUR PRESENTED METHOD IN THE AVEC2014 DATASET

Model Setting	RMSE	MAE
LA-CNN (CMP-)	8.72	7.16
LA-CNN (CMP+)	8.70	7.11
A2: LA-CNN (CMP-)+WSPP	8.51	6.82
B2: LA-CNN (CMP-)+WSPP	8.34	6.51
GA-CNN	9.02	7.41
C2: GA-CNN+WSPP	8.91	7.32
D2: LA-CNN(CMP-)+GA-CNN	8.62	6.91
E2: LA-CNN(CMP+)+GA-CNN	8.57	6.82
F2: LA-CNN(CMP-)+GA-CNN+WSPP	8.48	6.70
G2: LA-CNN(CMP+)+GA-CNN+WSPP	8.19	6.42

TABLE IV. THE RESULT OF THE RATE OF LEARNING IN THE EFFICIENCY OF OUR PRESENTED METHOD BY USING THE AVEC 2013 DATASET IN THE $G1$ MODE

Learning Rate	RMSE	MAE
0.000001	8.28	6.48
0.00001	8.50	6.79
0.0001	8.69	7.11
0.001	9.01	7.38
0.01	9.18	7.42
0.1	9.27	7.54

TABLE V. THE RESULT OF THE RATE OF LEARNING IN THE EFFICIENCY OF OUR PRESENTED METHOD BY USING THE AVEC 2014 DATASET IN THE $G2$ MODE

Learning Rate	RMSE	MAE
0.000001	8.18	6.40
0.00001	8.46	6.78
0.0001	8.71	7.13
0.001	8.94	7.37
0.01	9.12	7.39
0.1	9.26	7.55

In the following, to prove the efficiency of our presented approach, the comparison is made between the existing methods and our presented approach. It should be kept in mind that, as mentioned, the outcomes of our presented approach to compare the existing methods are provided on $G1$ mode and $G2$ mode. The quantitative performance comparison results for the dataset of AVEC 2013 and the dataset of AVEC 2014 are presented in Tables VI and VII. In particular, the presented models in [35]–[39] are based on the representations of the hand-crafted. Our approach performs better than the other approach in terms of the features of the hand-crafted emphasis on the experiences of the researchers, and it is hard to describe fully the depressive symptoms. In the approaches with the use of the DCNN, the presented method in [40] trains the deep methods in the big dataset, and next, it fine-tunes the dataset of AVEC 2013 and the dataset of AVEC 2014.

As shown in Tables VI and VII, our approach obtains the foremost efficiency between the approaches of the end-to-end in the used datasets. The presented method in [40] proposes a model of CNN based on the visual for depression detection by dividing roughly the area of the face into 3 parts and, next, by combining the total image of the face to better the model's detection performance. Our superior efficiency is according to the combination of the mechanism of the attention of the local and the mechanism of the attention of the global to extract the depression features. The results of the presented method in [40] display that their approach relies on the attention to just an area and also relinquishes other details of the face, which helps in depression detection. Reciprocally, the presented method in [41] obtains the acceptable efficiency sans a pre-trained network. The researchers segment the face area based on the points of the facial feature, and then, they clog the map of the feature to extract the information of the feature of the local. Our presented approach performs superior over these approaches with a significant margin.

D. Discussion

The obtained results from the experiments show that the proposed hand-crafted features as well as the feature aggregation method do not obtain higher RMSEs in compared to the method proposed in the present paper. From these results, it can be concluded that the proposed method in AVEC2013 and AVEC2014 can automatically learn the local and global feature information of the face area and outperform the best advanced methods for depression diagnosis. This further shows the effectiveness of the proposed method for diagnosis and analysis of depression.

In compared to the obtained results by similar methods, our method has improved the accuracy of depression diagnosis. There is a potential reason that the two-stage framework (eg, pre-training, fine-tuning) can effectively use their advantage to detect depression. By comparing the obtained results from similar methods, the RMSEs do not exceed them, our method is trained from scratch to be an end-to-end design for depression detection. For AVEC2014, as shown, our method achieves the comparable results to most video-based depression detection methods on the test set. In addition, these results of different components are better than the results obtained in AVEC2013.

Finally, in order to have an intuitive view of the prediction of the depression score from the images of the face, the images of the visualized face are provided in Fig. 8. The 1-th column from Fig. 8 displays the basic images. 2-th column to 5-th columns show the different areas from the facial image. The heat map on the images of the face is the fireplace area that the approach has learned. Before the merging of the attention maps, the proposed method can refer to multiple locations simultaneously. The proposed approach, in particular, relies on the areas of the movement of the associated facial muscles with depression, like the eyes, the mouth and the eyebrows. However, it ignores the irrelevant areas.

TABLE VI. COMPARISON OF OUR PRESENTED APPROACH TO SIMILAR APPROACHES BASED ON THE DATASET OF AVEC 2013

Methods	MAE	RSME
LPQ in [11]	10.88	13.61
PHOG in [42]]	-	10.45
LPQ-TOP in [2]	8.22	10.27
MRLBP-TOP, DPFV in [21]	7.55	9.20
LSOGCP in [39]	6.91	9.17
2D-CNN in [40]	6.50	8.41
3D-CNN in [41]	6.83	8.46
Proposed Method (G1)	6.48	8.30

TABLE VII. COMPARISON OF OUR PRESENTED APPROACH TO SIMILAR APPROACHES BASED ON THE DATASET OF AVEC 2014

Methods	MAE	RSME
[36]LGBP-TOP in	8.86	10.86
LBP-TOP in [43]	7.08	8.91
MRLBP-TOP, DPFV in [21]	7.21	9.01
LSOGCP in [39]	7.19	9.10
ResNet-50 in [44]	7.13	8.23
2D-CNN in [40]	6.51	8.39
3D-CNN in [41]	6.78	8.42
Proposed Method (G2)	6.42	8.19

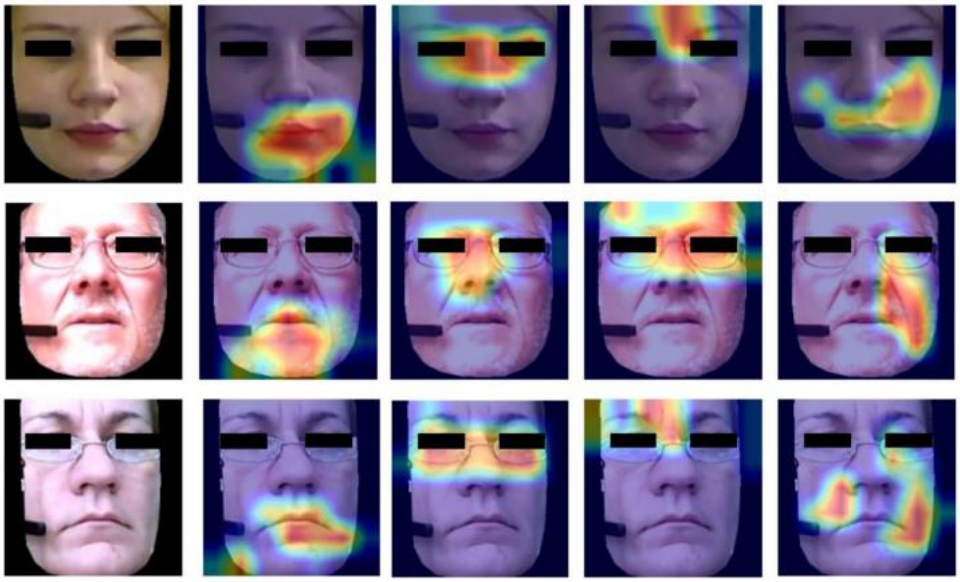

Fig. 8. The examples of the visualization of the face images with the different areas from the face.

V. CONCLUSIONS AND SUGGESTIONS

In this research, CNN, by the mechanism of attention, is used for the designing of an end-to-end integrated approach to the diagnosis of depression based on the video. We ratiocinate which a functional ability to take the feature pattern from the "encoded" depression on the areas of the face is important. In particular, a new framework is proposed. This framework consists of two branches: CNN based on local attention and CNN based on global attention. A CNN based on the local attention focuses only on the local patches. A CNN based on global attention learns the patterns of the global from the total area of the face. To take the information of the supplementary among 2 branches, a CNN basis on the local-global attention is presented. Finally, to achieve the informative patterns of the depression, a WSPP is applied to learn the final feature representations. The extensive tests in 2 datasets, namely AVEC2014 and AVEC2013, have displayed that the ability of our presented approach is higher than the diagnosis models of depression that are almost video-based.

Hereafter, the dataset from the further depressed patients will be gathered for the learning of the stronger features representation by the various appearance images. Additionally, the examination of learning of the multimodal depression representation (the audio, the video, the text, etc.) seems to be an interesting topic. In addition, we will investigate the more explainable patterns of the representation and the stronger approaches of the regression by the discriminant DCNN. Also, the presented model based on deep learning can aid doctors in the evaluation of depressed people.

ACKNOWLEDGMENT

This work was supported by the Hebei University of Water Resources and Electric Engineering Basic Research Project Funding: "Research on Social Recommendation Algorithms Based on Disentangled Graph Neural Networks" (No. SYKY2311).

REFERENCES

[1] S. Song, S. Jaiswal, L. Shen, M. Valstar, Spectral representation of behaviour primitives for depression analysis, IEEE Transactions on Affective Computing (2020), 1-1.

[2] L. Wen, X. Li, G. Guo, and Y. Zhu, “Automated depression diagnosis based on facial dynamic analysis and sparse coding,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 7, pp. 1432–1441, 2015.

[3] L. He, D. Jiang, and H. Sahli, “Multimodal depression recognition with dynamic visual and audio cues,” in 2015 International conference on affective computing and intelligent interaction (ACII), IEEE, 2015, pp. 260–266.

[4] A. Dhall and R. Goecke, “A temporally piece-wise fisher vector approach for depression analysis,” in 2015 International conference on affective computing and intelligent interaction (ACII), IEEE, 2015, pp. 255–259.

[5] S. Song, L. Shen, and M. Valstar, “Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral

features," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 158–165.

[6] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," IEEE Trans Affect Comput, vol. 11, no. 3, pp. 542–552, 2018.

[7] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[8] C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020), 1–1.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[10] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," arXiv preprint arXiv:1411.7923, 2014.

[11] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, F. Xu, 3d room layout estimation from a single rgb image, IEEE Transactions on Multimedia (2020), 1–1.

[12] M. Valstar et al., "Avec 2014: 3d dimensional affect and depression recognition challenge," in Proceedings of the 4th international workshop on audio/visual emotion challenge, 2014, pp. 3–10.

[13] 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, 2018.

[14] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in Proceedings of the 18th ACM international conference on multimodal interaction, 2016, pp. 445–450.

[15] Guo, W., Yang, H., Liu, Z., Xu, Y., and Hu, B. (2021). Deep neural networks for depression recognition based on 2d and 3d facial expressions under emotional stimulus tasks. Front. Neurosci. 15:609760. doi: 10.3389/fnins.2021.609760.

[16] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," Image Vis Comput, vol. 32, no. 10, pp. 641–647, 2014.

[17] Z. Liu, X. Yuan, Y. Li, Z. Shangguan, L. Zhou, and B. Hu, "PRA-Net: Part-and-Relation Attention Network for depression recognition from facial expression," Comput Biol Med, vol. 157, p. 106589, 2023.

[18] M. Niu, L. He, Y. Li, and B. Liu, "Depressioner: Facial dynamic representation for automatic depression level prediction," Expert Syst Appl, vol. 204, p. 117512, 2022.

[19] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," IEEE Trans Affect Comput, vol. 11, no. 3, pp. 542–552, 2018.

[20] 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, 2018.

[21] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," IEEE Trans Multimedia, vol. 21, no. 6, pp. 1476–1486, 2018.

[22] 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, 2017.

[23] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," IEEE Trans Cogn Dev Syst, vol. 10, no. 3, pp. 668–680, 2017.

[24] W. C. de Melo, E. Granger, and A. Hadid, "Combining global and local convolutional 3d networks for detecting depression from facial expressions," in 2019 14th ieee international conference on automatic face & gesture recognition (fg 2019), IEEE, 2019, pp. 1–8.

[25] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," IEEE Trans Affect Comput, vol. 13, no. 2, pp. 829–844, 2020.

[26] M. A. Uddin, J. B. Joolee, and Y.-K. Lee, "Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm," IEEE Trans Affect Comput, vol. 13, no. 2, pp. 864–870, 2020.

[27] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in 2016 IEEE winter conference on applications of computer vision (WACV), IEEE, 2016, pp. 1–10.

[28] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2983–2991.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[30] He, L., Tiwari, P., Lv, C., Wu, W., and Guo, L. (2022b). Reducing noisy annotations for depression estimation from facial images. Neural Netw. 153, 120–129. doi: 10.1016/j.neunet.2022.05.025.

[31] C. Yan, B. Gong, Y. Wei, and Y. Gao, "Deep multi-view enhancement hashing for image retrieval," IEEE Trans Pattern Anal Mach Intell, vol. 43, no. 4, pp. 1445–1451, 2020.

[32] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, and F. Xu, "3D room layout estimation from a single RGB image," IEEE Trans Multimedia, vol. 22, no. 11, pp. 3014–3024, 2020.

[33] J.B.J. Md Azher Uddin, Y.-K. Lee, Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm, IEEE Transactions on Affective Computing (2020), 1–1.

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv preprint arXiv:1711.05101, 2017.

[35] He, L., Guo, C., Tiwari, P., Su, R., Pandey, H. M., and Dang, W. (2022a). Depnet: an automated industrial intelligent system using deep learning for video-based depression analysis. Int. J. Intell. Syst. 37, 3815–3835. doi: 10.1002/int.22704.

[36] M. Valstar et al., "Avec 2014: 3d dimensional affect and depression recognition challenge," in Proceedings of the 4th international workshop on audio/visual emotion challenge, 2014, pp. 3–10.

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

[38] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," IEEE Trans Multimedia, vol. 21, no. 6, pp. 1476–1486, 2018.

[39] M. Niu, J. Tao, and B. Liu, "Local second-order gradient cross pattern for automatic depression detection," in 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, 2019, pp. 128–132.

[40] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," IEEE Trans Affect Comput, vol. 11, no. 3, pp. 542–552, 2018.

[41] L. He, C. Guo, P. Tiwari, H. M. Pandey, and W. Dang, "Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence," International Journal of Intelligent Systems, vol. 37, no. 12, pp. 10140–10156, 2022.

[42] Valstar, M., Schuller, B., Smith, K., Almaev, T. R., Eyben, F., Krajewski, J., et al. (2014). "AVEC 2014: 3D dimensional affect and depression recognition challenge," in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (Orlando, FL), 3–10. doi: 10.1145/2661806.2661807.

[43] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in 2015 International conference on affective computing and intelligent interaction (ACII), IEEE, 2015, pp. 255–259.

[44] W. C. De Melo, E. Granger, and A. Hadid, "Depression detection based on deep distribution learning," in 2019 IEEE international conference on image processing (ICIP), IEEE, 2019, pp. 4544–4548.