Spatial–Temporal Attention Network for Depression Recognition from Facial Videos

Article *in* Expert Systems with Applications · September 2023

DOI: 10.1016/j.eswa.2023.121410

7 authors, including:

Yuchen Pan
Capital Normal University
5 PUBLICATIONS **23** CITATIONS
SEE PROFILE

Tie Liu
Capital Normal University
36 PUBLICATIONS **3,954** CITATIONS
SEE PROFILE

Zhuhong Shao
Capital Normal University
57 PUBLICATIONS **1,049** CITATIONS
SEE PROFILE

Hui Ding
Capital Normal University
37 PUBLICATIONS **162** CITATIONS
SEE PROFILE

Spatial-Temporal Attention Network for Depression Recognition from Facial Videos

Yuchen Pan[a,1], Yuanyuan Shang[a,c,*,2], Tie Liu[a,d,3], Zhuhong Shao[a,d,4], Guodong Guo[b,5], Hui Ding[a,d,6] and Qiang Hu[e,7]

[a]College of Information Engineering, Capital Normal University, Beijing, 100048, China
[b]Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA
[c]Beijing Advanced Innovation Center for Imaging Technology, Beijing, 100048, China
[d]Beijing Key Laboratory of Electronic System Reliability Technology, Beijing, 100048, China
[e]Department of Psychiatry, ZhenJiang Mental Health Center, Zhenjiang, Jiangsu, 212000, China

ARTICLE INFO

Keywords:
Depression Recognition
Attention Mechanism
Video Recognition
Deep Learning
Visualization
Convolutional Neural Network

ABSTRACT

Recent studies focus on the utilization of deep learning approaches to recognize depression from facial videos. However, these approaches have been hindered by their limited performance, which can be attributed to the inadequate consideration of global spatial-temporal relationships in significant local regions within faces. In this paper, we propose Spatial-Temporal Attention Depression Recognition Network (STA-DRN) for depression recognition to enhance feature extraction and increase the relevance of depression recognition by capturing the global and local spatial-temporal information. Our proposed approach includes a novel Spatial-Temporal Attention (STA) mechanism, which generates spatial and temporal attention vectors to capture the global and local spatial-temporal relationships of features. To the best of our knowledge, this is the first attempt to incorporate pixel-wise STA mechanisms for depression recognition based on 3D video analysis. Additionally, we propose an attention vector-wise fusion strategy in the STA module, which combines information from both spatial and temporal domains. We then design the STA-DRN by stacking STA modules ResNet-style. The experimental results on AVEC 2013 and AVEC 2014 show that our method achieves competitive performance, with mean absolute error/root mean square error (MAE/RMSE) scores of 6.15/7.98 and 6.00/7.75, respectively. Moreover, visualization analysis demonstrates that the STA-DRN responds significantly in specific locations related to depression. The code is available at: https://github.com/divertingPan/STA-DRN.

1. Introduction

Major depressive disorder is characterized by persistent feelings of low mood and can affect individuals of all ages (Ackerman et al., 2018). Abnormal facial reactions during interviews with psychologists have been identified as an important diagnostic clue for depression (Fava & Kendler, 2000). This has inspired the increasing focus on recognizing depression based on facial videos, aiming to identify the relationship between depression levels and facial visual signals, and to recognize depression levels through algorithms that analyze facial videos. In previous studies (He et al., 2022; Zhu et al., 2018; de Melo et al., 2019a; Uddin et al., 2020; Jazaery & Guo, 2018; X. Zhou, Jin, et al., 2020), visual features in video signals have been extensively researched to identify potential patterns of depression and improve recognition performance.

Despite the increasing interest in utilizing facial videos to identify depression, there still exists a limited ability to extract features from significant local regions within the face. One crucial contextual element in video signals is the spatial characteristic of global facial actions. For example, a person might seem to be smiling with their mouth but have a calm or even furrowed brow, indicating a hesitant expression rather than a genuine smile. Another critical contextual factor is the temporal relationship among frames. While a person might exhibit distress during an unpleasant scene, it might not be enduring. However, individuals with depression tend to maintain a distressed facial expression, and this variability in facial expression is a dynamic process that necessitates capturing sequential global frames. Facial reactions related to depression are primarily reflected in the temporal and spatial changes of local facial areas, as shown in Fig. 1, but current

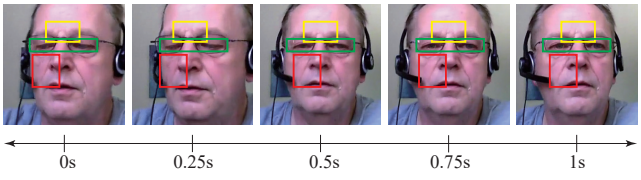


Figure 1: Facial reactions closely associated with depression are primarily manifested in temporal changes of spatial facial areas. For instance, persistent frowns are depicted by yellow boxes in the example, while dynamic eye movements are represented by green boxes. Dynamic facial features, such as changes in angle and illumination, are also taken into consideration when analyzing facial videos. The appearance of nasolabial folds, as shown in the red boxes, is notably influenced by shifts in facial direction.

*Corresponding author
2191002013@cnu.edu.cn (Y. Pan); yyshang@cnu.edu.cn (Y. Shang); 6578@cnu.edu.cn (T. Liu); zhshao@cnu.edu.cn (Z. Shao); Guodong.Guo@mail.wvu.edu (G. Guo); 5721@cnu.edu.cn (H. Ding); huqiang@sjtu.edu.cn (Q. Hu)

works rarely takes this aspect into consideration. Although video analysis has been developed (Yan & Woźniak, 2022), it is essential not only to locate the expression moment but also to model the correlation between multiple expressions, which could be extremely complex and implicit in a feature extractor. Deep learning models, rather than handcrafted extractors, are suitable for this task. Furthermore, unlike general video detection tasks such as facial detection (Wieczorek et al., 2022) or body detection (Woźniak et al., 2021), depression recognition typically uses pre-processed data with clear and full-frame facial images. Hence, related works focus on addressing the challenge of facial feature processing. Specifically, (Meng et al., 2013; Cummins et al., 2013; Zhu et al., 2018; X. Zhou, Jin, et al., 2020) use a single frame to recognize depression, resulting in the loss of dynamic facial expression information. In (Tran et al., 2015; Jazaery & Guo, 2018; de Melo et al., 2019a; X. Zhou, Wei, et al., 2020), a 3D convolutional neural network (CNN) (Tran et al., 2015) is used as the video feature extractor. However, the translation invariance of convolution and the down-sampling of pooling dilute the spatial-temporal information of the features and neglect the contributions from different spatial-temporal areas. As a result, the model takes all information equally, rather than concentrating on several temporal frames or spatial areas that are rich in essential information. Recent approaches (Niu et al., 2020; He et al., 2021; Niu et al., 2021) add weights to features with various attention mechanisms, boosting performance with features enhanced by adaptive weighted information. However, the spatial and temporal information is mixed during feature extraction, and the fusion strategy of spatial-temporal features fails to preserve the spatial-temporal structure.

In this paper, a Spatial-Temporal Attention Depression Recognition Network (STA-DRN) is proposed for depression recognition to enhance feature extraction and relevance by capturing both global and local spatial-temporal information. To achieve this, we first introduce spatial and temporal attention modules that employ a Spatial-Temporal Attention (STA) mechanism to correlate information between frames and pixels inspired by (Li et al., 2019; Woo et al., 2018). Then, we propose a fusion strategy that combines spatial and temporal attention vectors to yield a novel vector, which contains both local and global spatial-temporal information and the proposed STA module adaptively weights features. Furthermore, the STA-DRN is constructed by residual connecting (He et al., 2016) the STA modules.

Experiments are conducted on the Audio-Visual Emotion Challenge and Workshop (AVEC) 2013 (Valstar et al., 2013) and AVEC 2014 (Valstar et al., 2014) datasets, respectively, and demonstrate that our approach achieves a competitive result when compared to other vision-based approaches. Specifically, the mean absolute error (MAE) and root mean square error (RMSE) are 6.15/7.98 on AVEC 2013 and 6.00/7.75 on AVEC 2014, respectively. Furthermore, we utilize the modified Grad-CAM++ (Selvaraju et al., 2017) to illustrate that the model responds clearly and stably to multiple specific areas. This approach helps to reveal the relationship between facial action and depression level.

To summarize, our contributions are as follows:

1. We propose the STA mechanism to extract attention vectors with both global and local spatial-temporal information. This aspect has been overlooked in previous works. Furthermore, we develop a vector-wise fusion strategy to fuse spatial-temporal attention vectors in the STA module.
2. By incorporating the STA module with a popular deep learning network structure, we propose the STA-DRN. This approach can enhance the feature extraction and relevance in depression recognition using the STA mechanism.
3. Our experiments on the AVEC 2013 and AVEC 2014 datasets demonstrate the competitive performance of STA-DRN. Additionally, we apply visualization analysis to reveal the visual depression pattern from STA-DRN.

2. Related Work

Facial Depression Recognition. Prior research has demonstrated that vision-based approaches are proficient in recognizing depression. For instance, in AVEC 2013, the baseline approach utilized local phase quantization (Ojala et al., 2002) as the feature descriptor, and a support vector regression was trained on histograms from separated image blocks to predict the depression level. Meng *et al.*(Meng et al., 2013) proposed the Motion History Histogram, which records grayscale value changes for each pixel to describe motion information. They also extracted the Edge Orientation Histogram and Local Binary Patterns from the histogram to recognize depression. Additionally, Cummins *et al.*(Cummins et al., 2013) investigated the use of facial cues by combining the Space-Time Interest Points (Laptev et al., 2008) for detecting salient points and the Pyramid of Histogram of Gradients (Bosch et al., 2007) for reflecting the spatial-temporal changes.

These methods are primarily based on traditional handcrafted features. More recent works have focused on utilizing CNNs or a combination of CNNs and preprocessed handcrafted features. For instance, in (Zhu et al., 2018), an architecture employing two-stream CNNs was proposed to separately process facial images and optical flows. An end-to-end algorithm (X. Zhou, Jin, et al., 2020) exclusively utilized the image signal and divided it into different network pipelines. Melo *et al.*(de Melo et al., 2019b) utilized ResNet-50 as the backbone and proposed a method for transforming score regression into distribution prediction. Shang *et al.*(Shang et al., 2021) showed that local quaternion image features were effective when combined with the CNN model.

The aforementioned approaches employ either single frames or short-term dynamic features derived from consecutive frames (e.g., optical flow (Zhu et al., 2018) and quaternion local features (Shang et al., 2021)) for depression recognition. In other studies (Jazaery & Guo, 2018; de Melo

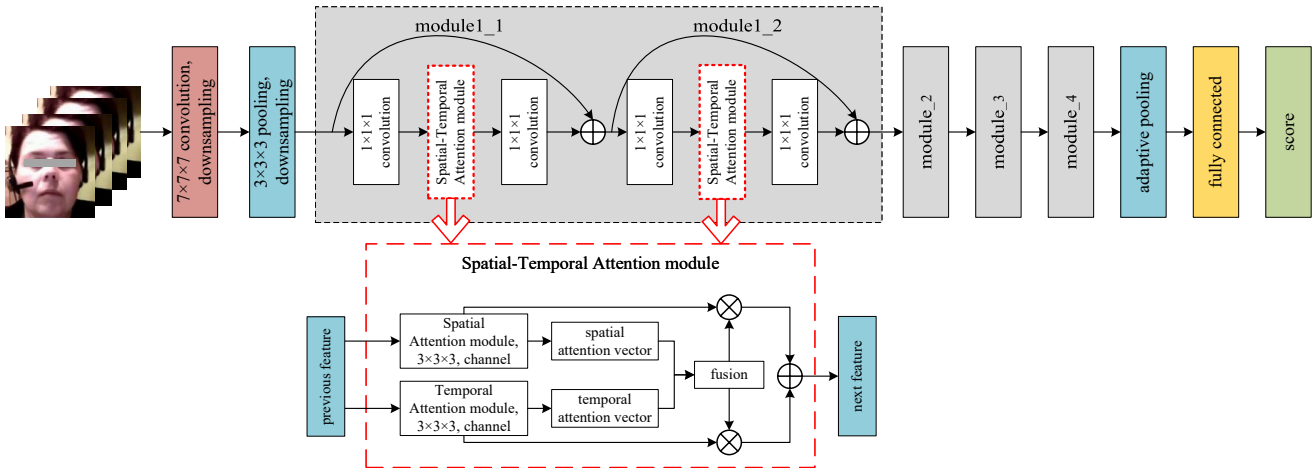


Figure 2: The overall architecture of STA-DRN. Table 1 shows more detailed parameters.

et al., 2019a; X. Zhou, Wei, et al., 2020), long-term video frames were directly extracted and predicted with 3D CNN. To effectively leverage the long-range temporal information, (Jazaery & Guo, 2018) extracted feature sequences from 3D convolution and employed RNN to predict the final score. (de Melo et al., 2019a) employed global and local 3D CNN that focused on the overall face and eye area without RNN structure. Lastly, (X. Zhou, Wei, et al., 2020) formulated depression score prediction as a label distribution learning problem and proposed a metric learning and 3D CNN-based approach.

Attention mechanism. The use of 3D methods is still limited in terms of effectively correlating global information and dynamically weighting features. To address this, attention mechanisms have been introduced in 2D image recognition tasks to enable adaptive weighted features. SENet (Hu et al., 2020), SKNet (Li et al., 2019) and ResNeSt (Zhang et al., 2020) all proposed different types of attention modules for this purpose, such as channel-wise attention and feature-map split attention. Woo *et al.*(Woo et al., 2018) proposed CBAM, which generates attention information across both channel and spatial dimensions. Similar adaptive weighting and partial information selecting techniques have been used in various vision tasks (Wu et al., 2023; Fernandez et al., 2019; Wu et al., 2021; Dosovitskiy et al., 2021). In video recognition, (X. Wang et al., 2018; Li et al., 2020) proposed non-local blocks to capture long-range dependencies along temporal axes, although this requires significant computation. For depression recognition, He *et al.*(He et al., 2021) proposed the Deep Local Global Attention Convolutional Neural Network, which can learn both global and local representations from facial images. Nonetheless, this approach may lack sufficient temporal information after feature extraction. In a multimodal approach proposed by Niu *et al.*(Niu et al., 2020), the spatial-temporal fusion in VSLF combines the spatial-temporal information of visual frames, but the feature representation is highly abstract, and the extraction of spatial attention information may interfere with positional relationships.

Visualization. Several CNN visualization techniques (B. Zhou et al., 2016; Selvaraju et al., 2017; Chattopadhay et al., 2018; H. Wang et al., 2020; Springenberg et al., 2015) have been developed to aid in understanding neural networks. In depression recognition research (X. Zhou, Jin, et al., 2020; Jazaery & Guo, 2018; X. Zhou, Wei, et al., 2020; de Melo et al., 2019b; Carneiro de Melo et al., 2020), these visualization methods have been utilized to display features related to depression. For example, in MR-DepressNet (X. Zhou, Jin, et al., 2020), stable responses in the eye areas were displayed through a DAM heatmap, which illustrated the depression-related features captured by the model. However, this visualization also revealed that the response was the only region highly related to MR-DepressNet. To strike a balance between speed and performance, we employed Grad-CAM++ (Chattopadhay et al., 2018) as our visualization method. We expanded the calculation dimension from 2D to 3D and optimized it for the depression recognition task to suit our model.

3. Proposed Method

This paper proposes STA-DRN, which aims to capture both spatial-temporal global and local information from video frames. To achieve this, the core of the STA modules uses the STA mechanism to enhance the relationship between information across pixels and frames. Specifically, the model incorporates a spatial attention module and a temporal attention module to extract spatial and temporal attention vectors, which can assign adaptive weights to features with spatial-temporal information. These two sub-modules are combined into an STA module with attention vector-wise fusion. In this section, we provide an overview of the overall structure of STA-DRN, followed by a detailed description of the spatial and temporal modules. Finally, we introduce the STA module, along with the fusion method.

3.1. Spatial-Temporal Attention Depression Recognition Network

Compared to previous 3D methods (Jazaery & Guo, 2018; de Melo et al., 2019a; X. Zhou, Wei, et al., 2020), our proposed STA-DRN not only incorporates temporal dynamic features, but also captures the relationship between spatial-temporal features through the STA module. The overall structure of the STA-DRN uses ResNet (He et al., 2016) as the backbone structure, which includes residual connections and bottleneck structures. The network architecture is illustrated in Fig. 2. In the first layer, a convolution operation with a kernel size of $7 \times 7 \times 7$ and a stride of $1 \times 2 \times 2$ is applied to extract and downsample low-level features. After a $3 \times 3 \times 3$ pooling layer with the stride of $1 \times 2 \times 2$, the features are then fed into a residual module containing two bottleneck structures. The proposed STA module replaces the middle layer of the bottleneck and generates a weighted feature with attention information. After a stack of residual modules, an adaptive pooling layer resamples the feature into a fixed shape. Finally, the last fully-connected layer predicts a score as the final output of the STA-DRN.

In the STA module, the spatial attention module and temporal attention module generate the inner features $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$, attention vector $\boldsymbol{s}_s$ and $\boldsymbol{s}_t$. The attention vectors are fused and then combined with the features, which is formulated as

$$\boldsymbol{X}_{fused} = (\boldsymbol{s}_s \odot \boldsymbol{s}_t) \odot (\boldsymbol{X}_s + \boldsymbol{X}_t), \tag{1}$$

where $\boldsymbol{X}_{fused}$ denotes the output of the STA module. More details are presented in Section 3.2.

3.2. Spatial-Temporal Attention Module

3.2.1. Spatial Attention Module

In video recognition, objects of interest often appear in a series of contiguous frames, and the ability to discern sequential information from features is essential for accurate recognition. Such information can be extracted from certain relative patterns that remain stable and spatially invariant across frames (X. Wang et al., 2018). Moreover, owing to the positional relationships between different facial features, the multiple spatial appearances of distinct patterns are also an important indicator for prediction, as depicted in Fig. 1. Consequently, we generate a spatial attention vector to enhance the inter-spatial relationships among features. To more effectively capture spatial information, CBAM (Woo et al., 2018) proposed a 2D spatial attention module. However, this module was developed for single-frame image recognition and the resulting spatial attention vector is generated across the entire channel axis without incorporating temporal information. Therefore, we propose a 3D video spatial attention module that draws inspiration from spatial information representation, as shown in Fig. 3. During feature extraction, we group features into distinct sets to increase the number of pattern combinations of feature extractors. In contrast to CBAM's spatial attention mechanism, we use soft-attention (Xu et al., 2015) to extract and combine spatial attention vectors from the spatial statistical information of features,

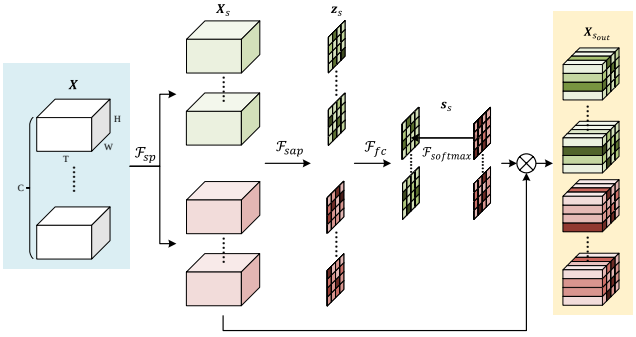


Figure 3: A diagram of the spatial attention module. The output feature $\boldsymbol{X}_{s_{out}}$ is generated by multiplying the attention vector $\boldsymbol{s}_s$ with the extracted feature $\boldsymbol{X}_s$.

thus fully utilizing global and local spatial information to generate the weights of each feature. With the soft-attention mechanism, the spatial attention vector can be considered as the specific global spatial activation of features obtained from convolutional local feature extractors.

Given the feature $\boldsymbol{X} \in \mathbb{R}^{T \times H \times W \times C}$ as the input of this module; T is the temporal length of the feature; $(H; W)$ is the spatial size of the feature and C is the number of channels. The function $\mathcal{F}_{sp}$ is a composite operation to get K feature groups $\boldsymbol{X}_s$ as:

$$\boldsymbol{X}_s = \mathcal{F}_{sp}(\boldsymbol{X}) = ReLU(BN(\mathcal{F}(\boldsymbol{X}))), \tag{2}$$

where $ReLU$ denotes the Rectified Linear Unit (Nair & Hinton, 2010) and BN denotes the Batch Normalization (Ioffe & Szegedy, 2015). The $\boldsymbol{X}_s \in \mathbb{R}^{T \times H \times W \times C' \times K}$; C' is the size of channels from an output of the local feature extractor, which is a convolution operator $\mathcal{F}$ with kernel size of $3 \times 3 \times 3$, and K denotes the split groups.

After that, the spatial statistics can be calculated from the feature $\boldsymbol{X}_s$. The 2D spatial feature statistical vector $\boldsymbol{z}_s \in \mathbb{R}^{H \times W \times C' \times K}$ can be computed with an average pooling operation $\mathcal{F}_{sap}$. Notice that the average pooling is operated on the temporal axis:

$$\boldsymbol{z}_s = \mathcal{F}_{sap}\left(\boldsymbol{X}_s\right) = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{X}_s(i). \tag{3}$$

The attention vector can be calculated from a fully connected operator $\mathcal{F}_{fc}$. In our practical implementation, the 1×1 convolution layer is recommended to replace the fully connected layer. Then the softmax operator is implemented in groups to combine the features and select the specific combination of patterns and features. The exact description of calculating the spatial attention vector is as follows:

$$\boldsymbol{s}_s = \mathcal{F}_{softmax}\left(\mathcal{F}_{fc}\left(\boldsymbol{z}_s\right)\right), \quad \boldsymbol{s}_s \in \mathbb{R}^{H \times W \times C' \times K}, \tag{4}$$

$$\mathcal{F}_{softmax}(\boldsymbol{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \quad j = 1 \ldots K. \tag{5}$$

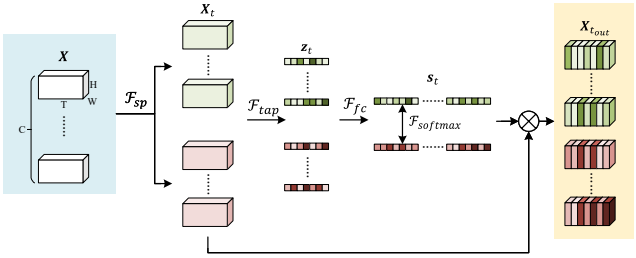


Figure 4: A diagram of the temporal attention module. The output feature $\boldsymbol{X}_{t_{out}}$ is generated by multiplying the attention vector $\boldsymbol{s}_t$ with the extracted feature $\boldsymbol{X}_t$.

The multiplication of attention vectors and extracted features is calculated along the T-axis of features with the Hadamard product to achieve the spatial attention weighted:

$$\boldsymbol{X}_{s_{out}}(t) = \boldsymbol{s}_s \odot \boldsymbol{X}_s(t), \quad t = 1 \dots T, \tag{6}$$

where $\boldsymbol{X}_{s_{out}} \in \mathbb{R}^{T\times H\times W\times(C'\times K)}$ is a 4D tensor that has a size of $(C' \times K)$ along the 4th axis.

3.2.2. Temporal Attention Module

Temporal variation within frames is a significant factor in video recognition. While short-term feature extraction can make use of dynamic information between a few frames, it is not suitable for extracting long-term dynamic features spanning several seconds. In the case of facial videos, both long-term and short-term features are equally important for recognition, as shown in Fig. 1. A previous method (Niu et al., 2020) employed an LSTM network to generate temporal information, but this approach is limited in its ability to capture long-term temporal relationships and can impede the parallelization of the deep model. In order to address these challenges, we propose a temporal attention model that employs a temporal attention vector to capture the long-term or short-term relationships between features. We introduce a temporal attention module to enhance temporal information, as illustrated in Fig. 4. Similar to the spatial attention module, the temporal attention vector is generated from temporal statistical information and attention mechanism to represent a long-term combination of short-term convolutional features.

Similar to the spatial module, the input feature $\boldsymbol{X}$ is split by $\mathcal{F}_{sp}$ which is the same as that shown in Eq. 2 in the spatial module to get K feature groups to generate $\boldsymbol{X}_t \in \mathbb{R}^{T\times H\times W\times C'\times K}$. Then the temporal statistical vector $\boldsymbol{z}_t \in \mathbb{R}^{T\times C'\times K}$ is calculated by the average pooling $\mathcal{F}_{tap}$:

$$\boldsymbol{z}_t = \mathcal{F}_{tap}\left(\boldsymbol{X}_t\right) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{X}_t(i,j). \tag{7}$$

After the fully connected $\mathcal{F}_{fc}$ and softmax $\mathcal{F}_{softmax}$, the temporal attention vector $\boldsymbol{s}_t$ is generated:

$$\boldsymbol{s}_t = \mathcal{F}_{softmax}\left(\mathcal{F}_{fc}\left(\boldsymbol{z}_t\right)\right), \quad \boldsymbol{s}_t \in \mathbb{R}^{T\times C'\times K}. \tag{8}$$

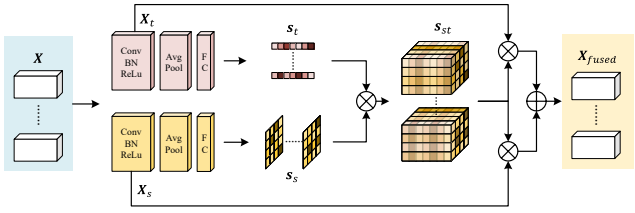


Figure 5: A diagram of attention vector-wise feature fusion method. The attention vector is utilized to capture the interaction of information between the spatial and temporal modules.

Finally, the temporal attention vector is utilized to weigh the extracted features along both the H-axis and W-axis of features through the Hadamard product, resulting in the generation of the weighted feature:

$$\boldsymbol{X}_{t_{out}}(h,w) = \boldsymbol{s}_t \odot \boldsymbol{X}_t(h,w), \quad h = 1 \dots H, w = 1 \dots W, \tag{9}$$

where $\boldsymbol{X}_{t_{out}} \in \mathbb{R}^{T\times H\times W\times(C'\times K)}$ is a 4D tensor that has a size of $(C' \times K)$ along the 4th axis.

3.2.3. Attention Vector-wise Feature Fusion

The conventional approach for combining spatial and temporal attention modules involves a simple summation of features from each branch of the model (Li et al., 2019). Formally, the fusion of spatial and temporal attention modules on a feature-wise basis can be represented as:

$$\boldsymbol{X}_{fused} = \boldsymbol{X}_{s_{out}} + \boldsymbol{X}_{t_{out}}, \tag{10}$$

where $\boldsymbol{X}_{s_{out}}$ and $\boldsymbol{X}_{t_{out}}$ denote the output features from spatial and temporal attention modules.

To exploit the attention information from spatial and temporal modules and incorporate the overall spatial-temporal information to the extracted feature, we propose a fusion method of attention vectors in the STA module. In the Non-local Neural Network (X. Wang et al., 2018), spatial-temporal information is produced by multiplying high-dimensional matrices, which consumes more memory during computation. Instead, we employ low-dimensional spatial and temporal attention vectors to achieve the integration of spatial-temporal information.

As shown in Fig. 5, the input feature is fed into temporal and spatial modules and two attention vectors are combined into a spatial-temporal attention vector with the following operation:

$$\boldsymbol{s}_{st}(t,h,w) = \boldsymbol{s}_t(t) \odot \boldsymbol{s}_s(h,w), \tag{11}$$

where $t = 1 \dots T, h = 1 \dots H, w = 1 \dots W$, $\boldsymbol{s}_t(t)$ and $\boldsymbol{s}_s(h,w)$ denotes the temporal and spatial attention vector, respectively.

Multiply the vector $\boldsymbol{s}_{st}$ with features from spatial module $\boldsymbol{X}_s$ and temporal module $\boldsymbol{X}_t$, then add the result to output the spatial-temporal feature:

$$\boldsymbol{X}_{fused} = \boldsymbol{s}_{st} \odot \left(\boldsymbol{X}_s + \boldsymbol{X}_t\right). \tag{12}$$

Table 1
The architecture for our STA-DRN. The module layer is the bottleneck block shown in Fig. 2. Downsampling is performed by module3_1, module4_1, and module5_1 with a stride of (1,2,2).

layer name	output size	layer
conv1	$64 \times 112 \times 112$	$7 \times 7 \times 7$, 64, stride (1,2,2)
module2_x	$64 \times 56 \times 56$	$3 \times 3 \times 3$ max pool, stride (1,2,2)
		$\begin{bmatrix} 1 \times 1 \times 1, 16 \\ 3 \times 3 \times 3, 16 \\ 1 \times 1 \times 1, 64 \end{bmatrix} \times 2$
module3_x	$64 \times 28 \times 28$	$\begin{bmatrix} 1 \times 1 \times 1, 32 \\ 3 \times 3 \times 3, 32 \\ 1 \times 1 \times 1, 128 \end{bmatrix} \times 2$
module4_x	$64 \times 14 \times 14$	$\begin{bmatrix} 1 \times 1 \times 1, 64 \\ 3 \times 3 \times 3, 64 \\ 1 \times 1 \times 1, 256 \end{bmatrix} \times 2$
module5_x	$64 \times 7 \times 7$	$\begin{bmatrix} 1 \times 1 \times 1, 128 \\ 3 \times 3 \times 3, 128 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 2$
fc1	$1 \times 1 \times 1$	adaptive average pool
		1-d fc

4. Experiments

In this section, we conduct experiments on the AVEC 2013 (Valstar et al., 2013) and AVEC 2014 (Valstar et al., 2014) datasets to evaluate the effectiveness of our proposed STA-DRN. We first perform an ablation study to assess the computational complexity and effectiveness of the STA-DRN, followed by visual analysis using the modified Grad-CAM++ (Chattopadhay et al., 2018) on the STA-DRN. Additionally, we evaluate the robustness of our STA-DRN to degraded inputs. Finally, we compare the performance of our proposed STA-DRN with other vision-based methods.

4.1. Datasets

We employ the AVEC 2013 and AVEC 2014 datasets for the training and evaluation of our proposed STA-DRN model. These datasets consist of videos accompanied by BDI-II score labels, which are self-evaluated by the participants in each video. The labels span from 0 to 63 and are divided into four depression levels: minimal (0-13), mild (14-19), moderate (20-28), and severe (29-63).

The AVEC 2013 dataset includes a total of 150 video clips, with each video exclusively featuring one individual. This dataset encompasses a total of 82 unique human participants. It is evenly divided into three separate subsets for training, development, and testing. Each video is presented in a colored MP4 format, with a resolution of 640×480 pixels and a frame rate of 30 frames per second. The age range of the participants spans from 18 to 63 years, with an average age of 31.5 years.

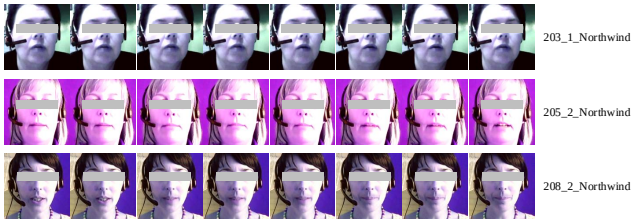


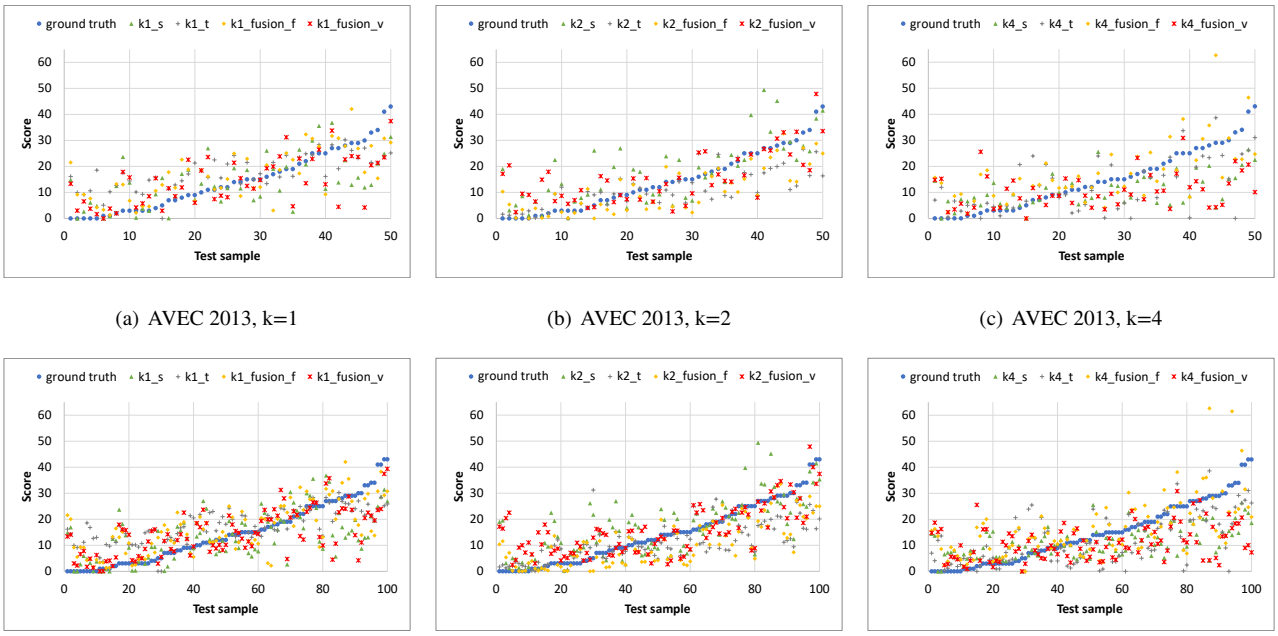
Figure 6: Some examples of input video frames. The augment is applied as the same set of parameters on one video section.

The AVEC 2014 dataset consists of a total of 300 video recordings, which are divided into three distinct sets: training, development, and testing sets. Each set contains two different types of video recordings: *Freeform* and *Northwind*. The *Northwind* task involves participants reading aloud the fable "Die Sonne und der Wind" (The North Wind and the Sun) in German. On the other hand, the *Freeform* task requires participants to answer a series of questions in the German language.

To maximize the preservation of short-term facial expressions, we extract video frames as images at a 1-frame interval. Our pre-processing phase utilizes the Dlib toolkit to extract facial landmarks, which serves to exclude background interference and align human faces. During alignment, we align the centers between the eyes and set the vertical distance between the eyes and mouth to be 1/3 of the image height. The aligned facial images are then rescaled to a size of 224×224 pixels. For training, we randomly select a sequence of 64 frames from a given video at a stochastic position to create a training clip. To augment the input data during training, we apply random horizontal flips and jitter on brightness, contrast, saturation, and hue ranging between 0 and 0.1 for all frames in one video clip, as illustrated in Figure 6. In the testing phase, we use the cropped and aligned video frames without any augmentation. For a given testing video, we crop it into a sub-video with 64 frames and predict each clip to calculate the mean value of depression scores, which in turn yield the final result for that video. We train our model on the union set of the AVEC 2013 and AVEC 2014 training sets, and subsequently validate and test our model on the development and testing sets of both datasets, respectively.

4.2. Training Details

Our model is implemented in the PyTorch environment (https://pytorch.org) on a server equipped with 6 TITAN RTX GPUs. The initial weights of the STA-DRN are pre-trained parameters on UCF101 (Soomro et al., 2012) for the temporal prior of videos, and on CK+ (Lucey et al., 2010) for the spatial prior of facial images. During the fine-tuning process, we set the training epoch to 200 and the batch size to 5. To prevent overfitting, we save the weights of the network when the validation losses reach a historical minimum. The network parameters are described in detail in Table 1. We use the Adam optimizer with an initial learning rate of 0.001. For



(a) AVEC 2013, k=1 (b) AVEC 2013, k=2 (c) AVEC 2013, k=4

(d) AVEC 2014, k=1 (e) AVEC 2014, k=2 (f) AVEC 2014, k=4

Figure 7: The visualization of the prediction vs ground truth with different attention modules on AVEC 2013 and 2014 test sets.

the learning rate schedule, we employ the cosine annealing warm restart:

$$\eta_t = \frac{1}{2} \times 0.001 \left(1 + \cos\left(\frac{T_{cur}}{T_i}\pi\right)\right). \tag{13}$$

T_{cur} represents the current epoch index, and T_i is the interval between two restart. The first restart iteration step $T_0 = 50$ and the factor increases after a restart is set to 2 according to the suggestion in (Loshchilov & Hutter, 2016).

The mean absolute error (MAE) and the root mean square error (RMSE) are used as evaluation functions to compare with other methods. The definitions of MAE and RMSE are as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right|, \tag{14}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2}, \tag{15}$$

where N denotes the number of samples, $\hat{y}_i$ represents the predicted result, and y_i denotes the true label of the i-th sample.

4.3. Results and Analysis

4.3.1. Analysis of STA Mechanism

To evaluate the effectiveness of STA mechanism, we incorporated different attention modules into the STA-DRN, namely the spatial attention module (*s*), temporal attention module (*t*), feature-wise fusion module (*fusion_f*), and attention vector-wise fusion module (*fusion_v*), to serve as the feature extractor. The experiments are performed on AVEC 2013 and AVEC 2014 test sets, and the results are presented in Table 2. In the table, k represents the number of split groups K in $\mathcal{F}_{sp}$ in Eq. 2, and *s/t/fusion_f/fusion_v* denotes

Table 2
Comparison of different structures of STA-DRN on AVEC 2013 and AVEC 2014. The suffix k is split groups, *s/t/fusion_f/fusion_v* denotes the embedded module of *spatial/temporal/feature-wise fusion/attention vector-wise fusion*. All other parameters are kept consistent with those shown in Table 1.

	AVEC 2013		AVEC 2014	
	MAE	RMSE	MAE	RMSE
k1_s	7.97	10.35	7.81	10.25
k1_t	7.58	9.12	7.68	9.29
k1_fusion_f	7.54	9.88	7.43	9.79
k1_fusion_v	7.19	9.65	7.07	9.55
k2_s	7.45	9.37	7.30	9.28
k2_t	7.21	9.32	7.07	9.22
k2_fusion_f	7.11	9.21	6.97	9.12
k2_fusion_v	**6.15**	**7.98**	**6.00**	**7.75**
k4_s	7.97	10.11	7.90	9.88
k4_t	7.77	10.12	7.48	9.56
k4_fusion_f	7.60	9.70	7.45	9.61
k4_fusion_v	7.41	9.83	7.38	9.63

Table 3
PCC and CCC of different STA-DRN structures on AVEC 2013 and AVEC 2014.

	AVEC 2013		AVEC 2014	
	PCC	CCC	PCC	CCC
k1_s	0.50	0.48	0.52	0.50
k1_t	0.65	0.52	0.66	0.52
k1_fusion_f	0.57	0.54	0.58	0.56
k1_fusion_v	0.59	0.57	0.60	0.59
k2_s	0.66	0.65	0.67	0.66
k2_t	0.67	0.56	0.68	0.58
k2_fusion_f	0.70	0.62	0.71	0.63
k2_fusion_v	**0.73**	**0.72**	**0.75**	**0.73**
k4_s	0.58	0.44	0.61	0.45
k4_t	0.58	0.57	0.63	0.59
k4_fusion_f	0.67	0.66	0.65	0.65
k4_fusion_v	0.67	0.68	0.67	0.67

the attention module used in STA-DRN. Furthermore, Figure 7 displays the predictions made by each model and the corresponding ground truth values for the AVEC 2013 and AVEC 2014 datasets.

The results of the experiments show that the models with *k2* split groups outperform those with *k1* split groups, demonstrating the superiority of the split feature group. However, increasing the number of split groups to *k4* leads to a negative performance, which suggests that a higher model complexity can result in optimization difficulty and overfitting effects. Comparing the performance of module fusion to that of a single spatial or temporal module, it is clear that the former is better (*fusion_f/fusion_v* vs. *s/t*). Additionally, the attention vector-wise fusion module (*fusion_v*) outperforms all the other modules, indicating that the spatial-temporal fusion module can improve performance more than a single temporal or spatial feature. This finding also highlights the superiority of the STA module over the feature-wise fusion module.

In addition, the Pearson Correlation Coefficient (PCC) and the Concordance Correlation Coefficient (CCC) are calculated using the following formulas:

$$\mathrm{PCC} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}, \tag{16}$$

$$\mathrm{CCC} = \frac{2\rho_{x,y}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + \left(\mu_x - \mu_y\right)^2}, \tag{17}$$

where cov denotes the covariance, σ_x and σ_y are the standard deviations. μ_x and μ_y are the mean values, and ρ denotes the PCC. The PCC and CCC of each variant model are listed in Table 3. The attention vector-wise fusion module with two split feature groups (*k2_fusion_v*) achieves the best performance with high PCC and CCC on both AVEC 2013 and AVEC 2014, which demonstrates a positive correlation between the predicted values and ground truth.

Table 4
Comparison of the performance, number of parameters, and FLOPs of different 3D approaches. P. and F. represent the number of parameters ($\times 10^6$) and FLOPs ($\times 10^9$), respectively.

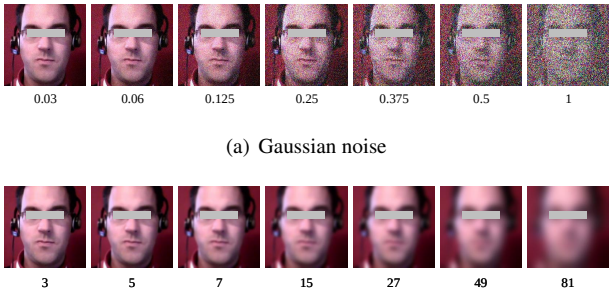
Model	AVEC 2013		AVEC 2014		P.	F.
	MAE	RMSE	MAE	RMSE		
ResNet-18	6.94	9.14	6.72	8.93	33.20	65.80
MDN-18	7.02	8.58	6.48	8.64	11.41	179.84
STA-18	**6.15**	7.98	**6.00**	**7.75**	31.27	192.71
ResNet-50	6.83	8.84	6.53	8.53	46.20	79.96
MDN-50	6.28	8.02	6.38	8.14	18.73	213.37
STA-50	6.39	8.19	6.27	7.94	41.41	238.16
ResNet-152	6.57	8.42	6.64	8.29	117.41	148.41
MDN-152	6.42	**7.83**	6.28	7.85	52.25	424.96
STA-152	6.48	8.16	6.64	8.25	78.14	375.69

4.3.2. Analysis of Computational Complexity

The performance and computational complexity of STA-DRN is compared with ResNet (He et al., 2016) and MDN (Carneiro de Melo et al., 2021) across different complexity levels. The experimental results, including the number of parameters (P.) and FLOPs (F.), are presented in Table 4. To ensure fair comparison, we use the same training and testing procedures and setups as described in Sections 4.1 and 4.2. The STA-18, STA-50, and STA-152 architectures are designed based on the specifications in Table 1, with bottleneck block numbers of $[2, 2, 2, 2]$, $[3, 4, 6, 3]$, and $[3, 8, 36, 3]$, respectively. Interestingly, the lighter STA-DRN (STA-18) outperforms its deeper counterparts (STA-50, STA-152) in terms of performance. Moreover, the more complex models exhibit no significant performance gains and may even lead to over-fitting. Remarkably, the STA-18 even outperforms more complicated models such as ResNet-50, MDN-50, and ResNet-152. Although MDN-152 yields a better performance in terms of AVEC 2013's RMSE, it requires significantly higher computational resources (424.96G FLOPs) compared to STA-18 (192.71G FLOPs).

4.3.3. Analysis of Degraded Inputs

We evaluate the robustness of STA-DRN to degraded inputs as shown in Fig. 8. The Gaussian noise $\mathcal{N} \sim (0, \sigma^2)$ with varying levels of standard deviation $\sigma = [0.03, 0.06, 0.125, 0.25, 0.375, 0.5, 1]$ is added to the original inputs, and the performance is evaluated as shown in Fig. 9. The baseline model is the optimal STA-DRN weights (k=2, STA-18), and the input images are tested directly. It shows that STA-DRN is robust to small amounts of noise ($\sigma < 0.25$), as the recognition error remains trivial. However, the recognition error increases dramatically when the noise level exceeds



(a) Gaussian noise

(b) Gaussian blur

Figure 8: The degraded input used to evaluate the robustness of STA-DRN. In (a), Gaussian noise is added to the input frames at various levels, and the numbers beneath indicate the σ of the noise. In (b), the input frames are Gaussian blurred using distinct kernel sizes, and the numbers beneath the images represent the kernel size.

0.375, and the trend shows that the error increases as the noise level increases.

Then the STA-DRN is tested with inputs under Gaussian blur with varying kernel sizes [3, 5, 7, 15, 27, 49, 81], and the results are shown in Fig. 10. As the kernel size increases, indicating blurrier inputs, the recognition error increases. This is expected, as the key facial features are lost in blurry images. However, the error rate rises more slowly when the kernel size exceeds 15, as the model tends to output an average recognition score with smoother and more average inputs.

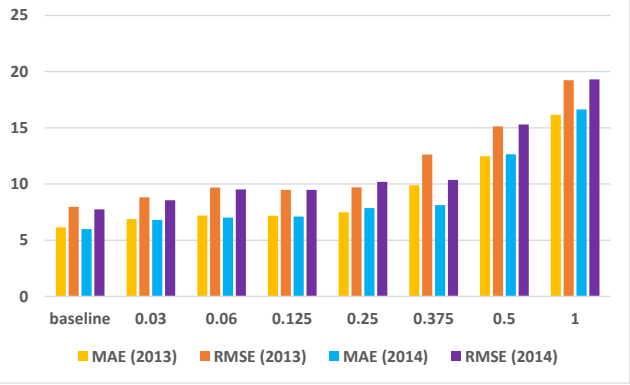


Figure 9: The comparison of STA-DRN performance in different noisy input levels on AVEC datasets.

4.3.4. Visualization Analysis

After evaluating various existing methods, such as Grad-CAM (Selvaraju et al., 2017), Grad-CAM++ (Chattopadhay et al., 2018), and Score-CAM (H. Wang et al., 2020), we selected Grad-CAM++ to measure the importance of each area towards the overall prediction. This method provides smoother and more stable heatmaps than Grad-CAM while being faster than Score-CAM.

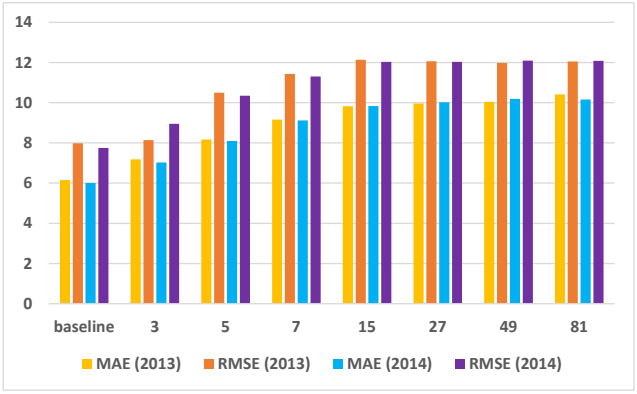


Figure 10: The comparison of STA-DRN performance in different blurring input levels on AVEC datasets.

To apply the Grad-CAM++ method to our depression recognition task, we made modifications to the calculation of feature weights. Specifically, we took the final prediction from the node of the last layer into account to better emphasize the effect of variant scores. The modified Grad-CAM++ generates the heatmap using the following approach:

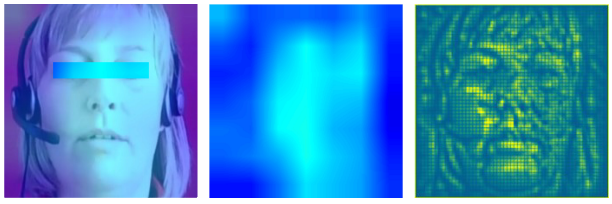
$$L = ReLU\left(\sum_k \omega\alpha_k A^k\right), \tag{18}$$

where ω denotes the output score of the model. A^k and α_k denote the k-th feature map and the corresponding weight which was presented in detail in (Chattopadhay et al., 2018). L responds to the importance of the highlighted area in the prediction. The area is related to the depression level.

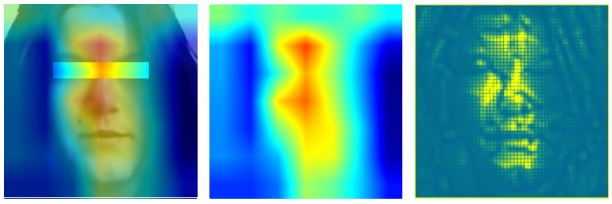
To provide visual examples, Fig. 11 shows some sample heatmaps overlaid on facial images (left column), corresponding heatmaps (middle column), and guided backpropagation (Springenberg et al., 2015) maps (right column) using the modified Grad-CAM++ method. To ensure clarity, we have sampled the image from the middle of a dynamic image series, and recolored the guided backpropagation map. Unlike the MR-DepressNet proposed in (X. Zhou, Jin, et al., 2020), our STA-DRN shows interesting areas that cover a range of facial features, including frown, eyes, nasolabial folds, and jaw. This aligns with the expectation that facial appearances related to depression detection distribute across these areas. Specifically, we observe that:

1. Input samples with low depression scores tend to exhibit no specific or minimally activated features across the entire facial area, as depicted in Fig. 11(a) and (b). However, as the depression score increases, certain areas become more intensively activated, as demonstrated in Fig. 11(c), (d), (e).
2. In the case of depression patients (samples with mild or higher levels), the highlighted features emerge in individual or combined areas, including the forehead, eye socket, nasolabial folds, cheek, and angulus oris. These areas exhibit a strong correlation with facial

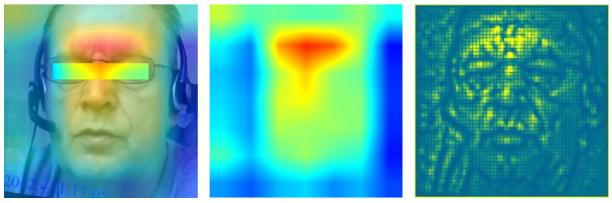
(a) ground truth: 0, prediction: 0.1374



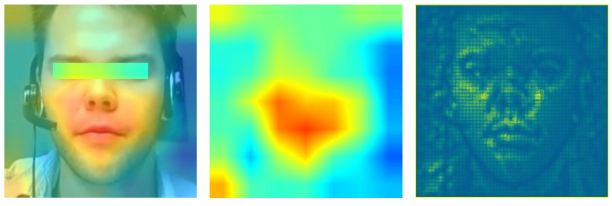
(b) ground truth: 3, prediction: 1.6536



(c) ground truth: 17, prediction: 15.9883



(d) ground truth: 25, prediction: 26.2520



(e) ground truth: 44, prediction: 43.5324

Figure 11: Some examples of the visualization with heatmaps (left column: overlay heatmaps on images, middle column: original heatmaps) and guided backpropagation maps (right column). The red areas in the heatmaps signify higher contribution to the final prediction, whereas the blue areas indicate lower contribution, providing clues to recognize depression. The guided backpropagation maps further highlight areas of high activation in yellow and low activation in cyan.

actions such as frowning and mouth curling. The regions associated with the most influential features are marked in red.

Table 5
Comparison with fine-tuned performance of popular pre-trained facial recognition models

Model	AVEC 2013		AVEC 2014	
	MAE	RMSE	MAE	RMSE
MobileFace (Deng et al., 2022)	8.15	9.34	8.04	9.20
IRSE50 (Hu et al., 2020)	8.34	9.86	8.25	9.67
FaceNet (Schroff et al., 2015)	7.05	8.64	7.04	8.51
IR152 (He et al., 2016)	6.57	8.42	6.64	8.29
STA-DRN (Ours)	**6.15**	**7.98**	**6.00**	**7.75**

3. The guided backpropagation maps illustrate the comprehensive features captured by our model, concentrating not just on the shape and contour of facial features, but also on potential textures and subtle expressions. In contrast to non-depressed individuals, whose maps present a "plain" feature across the entire area (Fig. 11(a) and (b)), individuals with depression guide features with greater contrast (Fig. 11(c), (d), (e)). This observation further substantiates the significance of distinct features in the context of depression recognition.

Our STA-DRN model produces reasonable predictions guided by specific features, as observed above. The heatmap generated by the model is consistent with the findings of MR-DepressNet (X. Zhou, Jin, et al., 2020), but the features identified by our model are more diverse. MR-DepressNet focuses primarily on the eyes, which is one of the regions of the multi-region network. This highlights the importance of eyes in facial depression recognition. However, the performance of MR-DepressNet is limited by its fixed focus regions. In contrast, the adaptive selection of spatial-temporal information in STA-DRN improves the diversity of identified features.

4.3.5. Comparison with Previous Works

First, we fine-tuned several popular facial recognition models, namely MobileFace (Deng et al., 2022), FaceNet (Schroff et al., 2015), IRSE50 (Hu et al., 2020), IR152 (He et al., 2016), and evaluated their performance on AVEC 2013 and AVEC 2014 testing sets. The results are presented in Table 5. Our analysis indicates that while pre-trained facial recognition models hold potential for fine-tuning and use in depression recognition, there remains a performance gap between these models and STA-DRN. One reason for this disparity is that these models rely on single-frame input, lacking the ability to extract temporal information. In contrast, STA-DRN incorporates spatial-temporal modules that enhance feature extraction and recognition.

The effectiveness of STA-DRN is demonstrated by testing it on the AVEC 2013 and AVEC 2014 testing sets and comparing it with previous work. The comparison is made with recent vision-based methods on AVEC 2013 dataset as shown in Table 6. The results indicate that the

Table 6
Comparison with previous vision-based methods on AVEC 2013 testing set

Method	MAE	RMSE
Baseline (Valstar et al., 2013)	10.88	13.61
PHOG (Cummins et al., 2013)	-	10.45
MHH + PLS (Meng et al., 2013)	9.14	11.19
LPQ (Kächele. et al., 2014)	8.97	10.82
LPQ-TOP (Wen et al., 2015)	8.22	10.27
DCNNs (Zhu et al., 2018)	7.58	9.82
DPFV (He et al., 2019)	7.55	9.20
RNN-C3D (Jazaery & Guo, 2018)	7.37	9.28
VSLF (Niu et al., 2020)	7.32	8.97
VLDN (Uddin et al., 2020)	7.04	8.93
DJ-LDML (X. Zhou, Wei, et al., 2020)	6.63	8.37
LGA-CNN-WSPP (He et al., 2021)	6.59	8.39
Global-Local C3D (de Melo et al., 2019a)	6.40	8.26
LQGDNet (Shang et al., 2021)	6.38	8.20
MR-DepressNet (X. Zhou, Jin, et al., 2020)	6.20	8.28
MDN (Carneiro de Melo et al., 2021)	6.24	**7.55**
STA-DRN (Ours)	**6.15**	7.98

Table 7
Comparison with previous vision-based methods on AVEC 2014 testing set

Method	MAE	RMSE
LGBP-TOP-SVM (Sidorov & Minker, 2014)	11.20	13.87
Baseline (Valstar et al., 2014)	8.86	10.86
INN (Cholet et al., 2019)	8.59	10.56
LBP-EOH-LPQ (Jan et al., 2014)	8.44	10.50
LPQ-DFT (Kaya et al., 2014)	8.20	10.27
DCNNs (Zhu et al., 2018)	7.47	9.55
RNN-C3D (Jazaery & Guo, 2018)	7.22	9.20
DPFV (He et al., 2019)	7.21	9.01
VLDN (Uddin et al., 2020)	6.86	8.78
DJ-LDML (X. Zhou, Wei, et al., 2020)	6.59	8.30
Global-Local C3D (de Melo et al., 2019a)	6.59	8.31
LGA-CNN-WSPP (He et al., 2021)	6.51	8.30
VSLF (Niu et al., 2020)	6.43	8.60
FDHH (Jan et al., 2018)	6.68	8.01
MR-DepressNet (X. Zhou, Jin, et al., 2020)	6.21	8.39
LQGDNet (Shang et al., 2021)	6.08	7.84
MDN (Carneiro de Melo et al., 2021)	6.06	**7.65**
STA-DRN (Ours)	**6.00**	7.75

MAE/RMSE of our method can reach 6.15/7.98. STA-DRN outperforms all the listed methods on MAE and reaches a competitive RMSE except for MDN, as reported by the authors (Carneiro de Melo et al., 2021). However, the number of parameters in STA-DRN is 31.27M, which is less than MDN (52.25M). Despite MDN's capacity to scale down to a model with 11.41M parameters, such downsizing significantly compromises its performance. While MDN's larger parameter configuration yields improved performance, the corresponding increase in computational complexity and model size is not justified by the performance gain, considering the substantial trade-off stemming from heightened complexity and size. In comparison with the CNN-based methods (X. Zhou, Wei, et al., 2020; Zhu et al., 2018; de Melo et al., 2019a; Shang et al., 2021; X. Zhou, Jin, et al., 2020) and the spatial-temporal methods (Uddin et al., 2020; Jazaery & Guo, 2018), our STA-DRN with an attention mechanism demonstrates superiority. Furthermore, STA-DRN achieves significantly better performance than the approaches with visual attention mechanism (He et al., 2021; Niu et al., 2020).

Upon comparing STA-DRN with the vision-based approaches on AVEC 2014, as presented in Table 7, STA-DRN achieves a MAE/RMSE of 6.00/7.75. It is evident that our STA-DRN outperforms the CNN (X. Zhou, Wei, et al., 2020; Zhu et al., 2018; de Melo et al., 2019a; Shang et al., 2021; X. Zhou, Jin, et al., 2020) and spatial-temporal (Uddin et al., 2020; Jazaery & Guo, 2018; Jan et al., 2018) approaches on AVEC 2014 in terms of MAE and ranks second in terms of RMSE. Furthermore, our method prevails over the Local-Global Attention method (He et al., 2021) in terms of incorporating temporal information into the attention mechanism. Moreover, our STA yields superior results when compared to the visual spatial-temporal attention VSLF (Niu et al., 2020), which also incorporates spatial-temporal information during feature extraction. It is important to note that the temporal vector in STA is generated using a simpler CNN structure, unlike VSLF, which employs both CNN and LSTM. Additionally, STA effectively preserves spatial information using a spatial attention vector, whereas VSLF employs a 3D CNN for spatial information that disrupts the spatial structure when the spatial feature is flattened.

5. Conclusion

We propose STA-DRN, a novel approach for recognizing depression based on sequential facial video frames, utilizing the STA mechanism. Firstly, we introduce a spatial and temporal attention module to effectively extract features by assigning adaptive weights, thereby enhancing the ability to capture both global and local information. Secondly, we propose a fusion strategy to combine the information from the spatial and temporal modules, via the STA module. Consequently, STA-DRN, comprised of the STA module, is developed to improve the feature extraction and relevance in depression recognition. Experiments on AVEC 2013 and AVEC 2014 validate the effectiveness of the proposed methodology, achieving competitive results with a MAE/RMSE of 6.15/7.98 on AVEC 2013 and 6.00/7.75 on

AVEC 2014, respectively. Moreover, our visualization analysis not only demonstrates a specific pattern obtained from STA-DRN, but also reveals the potential facial appearances that are indicative of varying levels of depression.

Although STA-DRN achieved satisfactory performance in recognizing depression using video data, it currently does not utilize audio signals. Incorporating speech features into the recognition model could potentially improve its performance, as has been demonstrated in recent multimodal approaches. Moreover, the AVEC datasets used to develop and test STA-DRN have limited visual signal data. Thus, collecting and maintaining a larger and more diverse visual dataset is critical for further improving the model's accuracy and developing a better depression recognition model.

In the future, we intend to develop a multimodal approach that incorporates not only video data but also audio, electroencephalogram (EEG), and magnetic resonance imaging (MRI), in order to fully utilize biological data. Our future research will focus on exploring joint feature representation of multimodal data, as well as constraints on multimodal features. Furthermore, we are interested in delving into the internal data flow of the multimodal deep learning network, with the objective of developing a novel visualization analysis.

CRediT authorship contribution statement

Yuchen Pan: Methodology, Software, Validation, Resources, Writing - Original Draft, Writing - Review & Editing. **Yuanyuan Shang:** Conceptualization, Validation, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. **Tie Liu:** Methodology, Validation, Writing - Review & Editing. **Zhuhong Shao:** Methodology, Validation, Writing - Review & Editing. **Guodong Guo:** Validation, Writing - Original Draft, Writing - Review & Editing. **Hui Ding:** Methodology, Validation, Writing - Review & Editing. **Qiang Hu:** Software, Validation, Writing - Review & Editing.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61876112, 61601311), and the Natural Science Foundation of Beijing (L201022).

References

Ackerman, I., Buchbinder, R., Chin, K., Cicuttini, F., Driscoll, T., Gall, S., … GBD 2017 Disease and Injury Incidence and Prevalence Collaborators (2018, November 10). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, *392*(10159), 1789–1858. doi: 10.1016/S0140-6736(18)32279-7

Bosch, A., Zisserman, A., & Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th acm international conference on image and video retrieval* (p. 401–408). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1282280.1282340 doi: 10.1145/1282280.1282340

Carneiro de Melo, W., Granger, E., & Hadid, A. (2020). A deep multiscale spatiotemporal network for assessing depression from facial dynamics. *IEEE Transactions on Affective Computing*, 1-1. doi: 10.1109/TAFFC.2020.3021755

Carneiro de Melo, W., Granger, E., & Bordallo Lopez, M. (2021). Mdn: A deep maximization-differentiation network for spatio-temporal depression detection. *IEEE Transactions on Affective Computing*, 1-1.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 ieee winter conference on applications of computer vision (wacv)* (p. 839-847). doi: 10.1109/WACV.2018.00097

Cholet, S., Paugam-Moisy, H., & Regis, S. (2019). Bidirectional associative memory for multimodal fusion: a depression evaluation case study. In *2019 international joint conference on neural networks (ijcnn)* (p. 1-6).

Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., & Epps, J. (2013). Diagnosis of depression by behavioural signals: A multimodal approach. In *Proceedings of the 3rd acm international workshop on audio/visual emotion challenge* (p. 11–20). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2512530.2512535 doi: 10.1145/2512530.2512535

de Melo, W. C., Granger, E., & Hadid, A. (2019a). 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)* (p. 1-8). doi: 10.1109/FG.2019.8756568

de Melo, W. C., Granger, E., & Hadid, A. (2019b). Depression detection based on deep distribution learning. In *2019 ieee international conference on image processing (icip)* (p. 4544-4548). doi: 10.1109/ICIP.2019.8803467

Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., & Zafeiriou, S. (2022). Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 5962-5979. doi: 10.1109/TPAMI.2021.3087709

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th international conference on learning representations, ICLR 2021, virtual event, austria, may 3-7, 2021*. OpenReview.net. Retrieved from https://openreview.net/forum?id=YicbFdNTTy

Fava, M., & Kendler, K. (2000, November). Major depressive disorder. *Neuron*, *28*(2), 335–341. Retrieved from https://doi.org/10.1016/s0896-6273(00)00112-4 doi: 10.1016/s0896-6273(00)00112-4

Fernandez, P. D. M., Peña, F. A. G., Ren, T. I., & Cunha, A. (2019). Feratt: Facial expression recognition with attention net. In *2019 ieee/cvf conference on computer vision and pattern recognition workshops (cvprw)* (p. 837-846). doi: 10.1109/CVPRW.2019.00112

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 ieee conference on computer vision and pattern recognition (cvpr)* (p. 770-778). doi: 10.1109/CVPR.2016.90

He, L., Chan, J. C.-W., & Wang, Z. (2021). Automatic depression recognition using cnn with attention mechanism from videos. *Neurocomputing*, *422*, 165-175. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925231220315101 doi: https://doi.org/10.1016/j.neucom.2020.10.015

He, L., Jiang, D., & Sahli, H. (2019). Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding. *IEEE Transactions on Multimedia*, *21*(6), 1476-1486.

He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., … Dang, W. (2022). Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, *80*, 56-86. Retrieved from https://www.sciencedirect.com/science/article/pii/S1566253521002207 doi: https://doi.org/10.1016/j.inffus.2021.10.012

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2020). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(8), 2011-2023. doi: 10.1109/TPAMI.2019.2913372

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep

network training by reducing internal covariate shift. In *Proceedings of the 32nd international conference on international conference on machine learning - volume 37* (p. 448–456). JMLR.org.

Jan, A., Meng, H., Gaus, Y. F. A., Zhang, F., & Turabzadeh, S. (2014). Automatic depression scale prediction using facial expression dynamics and regression. In *Proceedings of the 4th international workshop on audio/visual emotion challenge* (p. 73–80). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2661806.2661812 doi: 10.1145/2661806.2661812

Jan, A., Meng, H., Gaus, Y. F. B. A., & Zhang, F. (2018). Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(3), 668-680. doi: 10.1109/TCDS.2017.2721552

Jazaery, M. A., & Guo, G. (2018). Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing*, 1-1. doi: 10.1109/TAFFC.2018.2870884

Kaya, H., Çilli, F., & Salah, A. A. (2014). Ensemble cca for continuous emotion prediction. In *Proceedings of the 4th international workshop on audio/visual emotion challenge* (p. 19–26). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2661806.2661814 doi: 10.1145/2661806.2661814

Kächele., M., Glodek., M., Zharkov., D., Meudt., S., & Schwenker., F. (2014). Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. In *Proceedings of the 3rd international conference on pattern recognition applications and methods - icpram* (p. 671-678). SciTePress. doi: 10.5220/0004828606710678

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *2008 ieee conference on computer vision and pattern recognition* (p. 1-8). doi: 10.1109/CVPR.2008.4587756

Li, J., Liu, X., Zhang, M., & Wang, D. (2020). Spatio-temporal deformable 3d convnets with attention for action recognition. *Pattern Recognition*, *98*, 107037. Retrieved from http://www.sciencedirect.com/science/article/pii/S0031320319303383 doi: https://doi.org/10.1016/j.patcog.2019.107037

Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. In *2019 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 510-519). doi: 10.1109/CVPR.2019.00060

Loshchilov, I., & Hutter, F. (2016, August). SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv e-prints*, arXiv:1608.03983.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition - workshops* (p. 94-101). doi: 10.1109/CVPRW.2010.5543262

Meng, H., Huang, D., Wang, H., Yang, H., AI-Shuraifi, M., & Wang, Y. (2013). Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd acm international workshop on audio/visual emotion challenge* (p. 21–30). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2512530.2512532 doi: 10.1145/2512530.2512532

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on international conference on machine learning* (p. 807–814). Madison, WI, USA: Omnipress.

Niu, M., Liu, B., Tao, J., & Li, Q. (2021). A time-frequency channel attention and vectorization network for automatic depression level prediction. *Neurocomputing*, *450*, 208-218. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925231221005981 doi: https://doi.org/10.1016/j.neucom.2021.04.056

Niu, M., Tao, J., Liu, B., Huang, J., & Lian, Z. (2020). Multimodal spatiotemporal representation for automatic depression level detection. *IEEE Transactions on Affective Computing*, 1-1. doi: 10.1109/TAFFC.2020.3031345

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971-987. doi: 10.1109/TPAMI.2002.1017623

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (p. 815-823). doi: 10.1109/CVPR.2015.7298682

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 ieee international conference on computer vision (iccv)* (p. 618-626). doi: 10.1109/ICCV.2017.74

Shang, Y., Pan, Y., Jiang, X., Shao, Z., Guo, G., Liu, T., & Ding, H. (2021). Lqgdnet: A local quaternion and global deep network for facial depression recognition. *IEEE Transactions on Affective Computing*, 1-1. doi: 10.1109/TAFFC.2021.3139651

Sidorov, M., & Minker, W. (2014). Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach. In *Proceedings of the 4th international workshop on audio/visual emotion challenge* (p. 81–86). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2661806.2661816 doi: 10.1145/2661806.2661816

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, *abs/1212.0402*. Retrieved from http://arxiv.org/abs/1212.0402

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. A. (2015). Striving for simplicity: The all convolutional net. *CoRR*, *abs/1412.6806*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *2015 ieee international conference on computer vision (iccv)* (p. 4489-4497). doi: 10.1109/ICCV.2015.510

Uddin, M. A., Joolee, J. B., & Lee, Y. (2020). Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm. *IEEE Transactions on Affective Computing*, 1-1. doi: 10.1109/TAFFC.2020.2970418

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., ... Pantic, M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge* (p. 3–10). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2661806.2661807 doi: 10.1145/2661806.2661807

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., ... Pantic, M. (2013). Avec 2013: The continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd acm international workshop on audio/visual emotion challenge* (p. 3–10). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2512530.2512533 doi: 10.1145/2512530.2512533

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., ... Hu, X. (2020). Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 ieee/cvf conference on computer vision and pattern recognition workshops (cvprw)* (p. 111-119). doi: 10.1109/CVPRW50498.2020.00020

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *2018 ieee/cvf conference on computer vision and pattern recognition* (p. 7794-7803). doi: 10.1109/CVPR.2018.00813

Wen, L., Li, X., Guo, G., & Zhu, Y. (2015). Automated depression diagnosis based on facial dynamic analysis and sparse coding. *IEEE Transactions on Information Forensics and Security*, *10*(7), 1432-1441. doi: 10.1109/TIFS.2015.2414392

Wieczorek, M., Siłka, J., Woźniak, M., Garg, S., & Hassan, M. M. (2022). Lightweight convolutional neural network model for human face detection in risk situations. *IEEE Transactions on Industrial Informatics*, *18*(7), 4820-4829. doi: 10.1109/TII.2021.3129629

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision – eccv 2018* (pp. 3–19). Cham: Springer International Publishing.

Woźniak, M., Siłka, J., & Wieczorek, M. (2021). Deep learning based crowd counting model for drone assisted systems. In *Proceedings of the 4th acm mobicom workshop on drone assisted wireless communications for 5g and beyond* (p. 31–36). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3477090.3481054 doi: 10.1145/3477090.3481054

Wu, Z., Liu, C., Wen, J., Xu, Y., Yang, J., & Li, X. (2023). Selecting high-quality proposals for weakly supervised object detection with bottom-up aggregated attention and phase-aware loss. *IEEE Transactions on Image Processing*, *32*, 682-693. doi: 10.1109/TIP.2022.3231744

Wu, Z., Wen, J., Xu, Y., Yang, J., & Zhang, D. (2021). Multiple instance detection networks with adaptive instance refinement. *IEEE Transactions on Multimedia*, 1-1. doi: 10.1109/TMM.2021.3125130

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., … Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd international conference on international conference on machine learning - volume 37* (p. 2048–2057). JMLR.org.

Yan, G., & Woźniak, M. (2022, jun). Accurate key frame extraction algorithm of video action for aerobics online teaching. *Mob. Netw. Appl.*, *27*(3), 1252–1261. Retrieved from https://doi.org/10.1007/s11036-022-01939-1 doi: 10.1007/s11036-022-01939-1

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., … Smola, A. (2020, April). ResNeSt: Split-Attention Networks. *arXiv e-prints*, arXiv:2004.08955.

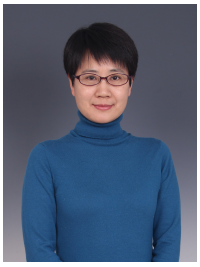
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *2016 ieee conference on computer vision and pattern recognition (cvpr)* (p. 2921-2929). doi: 10.1109/CVPR.2016.319

Zhou, X., Jin, K., Shang, Y., & Guo, G. (2020). Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*, *11*(3), 542-552. doi: 10.1109/TAFFC.2018.2828819

Zhou, X., Wei, Z., Xu, M., Qu, S., & Guo, G. (2020). Facial depression recognition by deep joint label distribution and metric learning. *IEEE Transactions on Affective Computing*, 1-1. doi: 10.1109/TAFFC.2020.3022732

Zhu, Y., Shang, Y., Shao, Z., & Guo, G. (2018). Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing*, *9*(4), 578-584. doi: 10.1109/TAFFC.2017.2650899

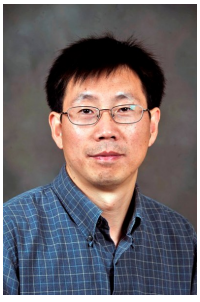
Yuchen Pan received his B.E. degree from the College of Computer Science and Technology, Harbin Engineering University in 2019, And received the M.S. degree from the College of Information Engineering, Capital Normal University in 2022. He is now with the Intelligent Recognition and Image Processing Laboratory of Capital Normal University, working on depression recognition in the field of computer vision and deep learning. He has authored papers in peer reviewed journals including IEEE Transactions on Affective Computing and licensed China patents. His research interest includes computer vision, digital image processing and deep learning.



Yuanyuan Shang received the Ph.D. degree from Chinese Academy of Sciences in 2005. She is currently a professor and vice dean of the Graduate School, Capital Normal University, Beijing, P.R. China. Her research interests include computer vision and medical image processing. She has authored more than 70 scientific papers in peer-reviewed journals and conferences, including some top venues such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Affective Computing, CVPR and ACM Multimedia. She is currently serving as vice president of Beijing Artificial Intelligent Society.

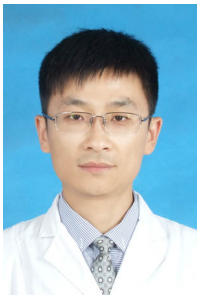
Tie Liu received the BS, MS, and PhD degrees from Xian Jiaotong University, in 2001, 2004, and 2007, respectively. He is currently an associate professor in the College of Information Engineering, Capital Normal University. His areas of interest include machine learning, pattern recognition, multimedia computing, and computer vision. He is also interested in data analysis and mining.

Zhuhong Shao received the B.S. degree in Biomedical Engineering from Jilin Medical University, Jilin, China, in 2009, and the M.S. degree in Electrical Engineering from Beijing Jiaotong University, Beijing China, in 2011 and the Ph.D. degree in Computer Science and Technology from Southeast University, Nanjing, China, in 2015. He is currently an associate professor in the College of Information Engineering, Capital Normal University, Beijing, China. His research interest includes computer vision and multimedia information security.



Guodong Guo received his B.E. degree in Automation from Tsinghua University, Beijing, China, in 1991, the Ph.D. degree in Pattern Recognition and Intelligent Control from Chinese Academy of Sciences, in 1998, and the Ph.D. degree in computer science from the University of WisconsinMadison, in 2006. He is currently an Associate Professor in the Lane Department of Computer Science and Electrical Engineering, West Virginia University. In the past, he visited and worked in several places, including INRIA, Sophia Antipolis, France, Ritsumeikan University, Japan, Microsoft Research, China, and North Carolina Central University. He won the North Carolina State Award for Excellence in Innovation in 2008, and Outstanding New Researcher of the Year (2010-2011) at CEMR, WVU. His research areas include computer vision, machine learning, and multimedia. He has authored a book, Face, Expression, and Iris Recognition Using Learning-based Approaches (2008), published over 60 technical papers in face, iris, expression, and gender recognition, age estimation, multimedia information retrieval, and image analysis, and filed three patents on iris and texture image analysis. He is an editorial board member of the IET Biometrics, and a senior member of IEEE.

Hui Ding received her Ph.D. degree from the School of Information Science and Technology, Beijing Institute of Technology of China, in 2006. She is currently an Associate Professor with the College of Information Engineering, Capital Normal University, Beijing, China. She has authored over 30 scientific papers in peer-reviewed journals and conferences. Her research interests include computer vision, medical image processing, and machine learning.



Qiang Hu received his Ph.D. degree from Shanghai Jiaotong University, in 2022. He is currently working at the Department of Psychiatry, Zhen-Jiang Mental Health Center, Zhenjiang, Jiangsu, China. His research interests include the conduct of systematic reviews and the treating of patients with psychoses.