Journal Pre-proof

Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer

Yongfeng Tao, Minqiang Yang, Yushan Wu, Kevin Lee, Adrienne Kline, Bin Hu

PII: S2352-8648(23)00063-9

DOI: https://doi.org/10.1016/j.dcan.2023.03.007

Reference: DCAN 644

To appear in: *Digital Communications and Networks*

Received Date: 12 December 2022

Revised Date: 15 February 2023

Accepted Date: 11 March 2023

Please cite this article as: Y. Tao, M. Yang, Y. Wu, K. Lee, A. Kline, B. Hu, Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer, *Digital Communications and Networks* (2023), doi: https://doi.org/10.1016/j.dcan.2023.03.007.

Digital Communications and Networks(DCN)

HOSTED BY

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/dcan

Depressive Semantic Awareness from Vlog Facial and Vocal Streams via Spatio-temporal Transformer

Yongfeng Tao[a], Minqiang Yang[*a], Yushan Wu[a], Kevin Lee[b], Adrienne Kline[c], Bin Hu[*a]

[a]School of Information Science and Engineering, Lanzhou University, Lanzhou, China
[b]The School of Accounting, Auditing and Taxation, Business School, UNSW Sydney
[c]Department of Preventive Medicine, Northwestern University, Chicago, IL, United States

Abstract

With the rapid rise of information transmission via the Internet, efforts have been made to reduce network load to promote efficiency. One such application is semantic computing, which can extract and process semantic communication. Social media has enabled users to share their current emotions, opinions, and life events through their mobile devices. Notably, people suffering from mental health problems are more willing to share their feelings on social networks. Therefore, it is necessary to extract semantic information from social media (vlog data) to identify abnormal emotional states to facilitate early identification and intervention. Most studies have not considered spatio-temporal information when fusing multimodal information to identify abnormal emotional states such as depression. To solve this problem, this paper proposes a spatio-temporal squeeze transformer method for the extraction of semantic features of depression. First, we embed the module with spatio-temporal data into the transformer encoder, which is leveraged to obtain spatio-temporal feature representations. Second, a classifier with a voting mechanism is designed to encourage the model to classify depression and non-depression effectively. Experiments are conducted on the D-Vlog dataset. The results show that the method is effective, and the accuracy rate can reach 70.70%. This work provides scaffolding for future work in the detection of affect recognition in semantic communication based on social media vlog data.

KEYWORDS: Emotional computing, Semantic awareness, Depression recognition, Vlog data

1. Introduction

Alongside the exponential expansion of Internet information transmission, wireless communication technology has developed iteratively, and has experienced five generations thus far [1]. However, with an exponentially growing global demand for data communications, current wireless communication technology are heavily taxed. Combine cloud and mobile edge computing to reduce execution latency and offload programme efficiency to mobile devices [2]. Scholars are predicting that the fifth generation (5G) will reach the service threshold within 20 years [3]. Semantic communication (SC) affords an opportunity to solve this problem. Weaver et al. [4] proposed a three-layer SC model. In the first layer, namely the technical layer, entails 'how to transmit information from the transmitter to the receiver by bits or symbols'; the second layer, namely the semantic layer, involves 'how to

*Minqiang Yang (Corresponding author) is with School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China (e-mail: yangmq@lzu.edu.cn). Bin Hu (Corresponding author) is with Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Gansu, China (e-mail: bh@lzu.edu.cn).

[1]Yongfeng Tao and Yushan Wu are with School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China (e-mail: taoyf21@lzu.edu.cn, wuysh2021@lzu.edu.cn). Kevin Lee is with the School of Accounting, Auditing and Taxation, Business School, UNSW Sydney (e-mail: k.li@unsw.edu.au). Adrienne Kline is with Department of Preventive Medicine, Northwestern University, Chicago, IL, United States (e-mail: adrienne.kline@northwestern.edu).

convey the desired meaning of bits or symbols'; the third layer, namely the effectiveness layer, involves 'how the receiver's behaviors are affected by the received meanings' [5][6]. The comparison of the first layer and the second layer communication policies allows semantic communication to more accurately extract the purpose of the transmission [7], remove semantically irrelevant/redundant information from the transmission [5], reducing the burden on networks. Moreover, the receiver can correct transmission error and recover the received information to build a more accurate transmission [7].

Depression, is the most prevalent mental disorder and affects more than 350 million people worldwide [8]. Its negative impact takes its toll on individuals through persistent negative emotions that may lead to self harm [9]. Additionally, depression poses a burden on society, the healthcare system and those in the individual's social network. It is expected to carry the highest global disease burden (GDB) by 2030 [10]. Yet, despite the severity and prevalence of depression, many patients are misdiagnosed and therefore left untreated. This is largely due to diagnostic methods for depression relying on an expert's subjective perception [11]. At present, behavioral and physiological signals are used to identify various diseases [12][13], including depression [14]. Taken together, using information from both modalities can explore true emotions/mood. Early on, researchers used a variety of hand-crafted features [15, 16] to capture depression information, but most of these methods failed to capture high-level semantics. In recent years, researchers have increasingly applied deep neural networks [17, 18] to depression detection with good results, but it remains a challenge to extract and fuse spatio-temporal information from the data stream more effectively.

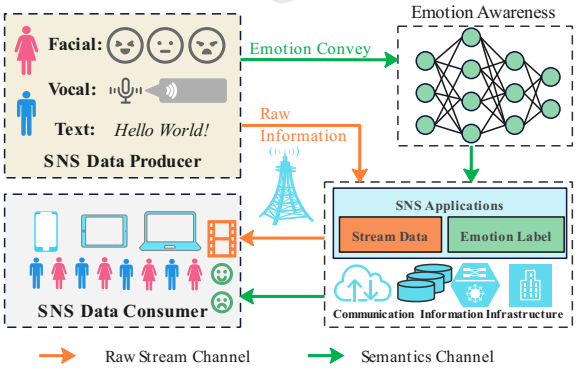


Fig. 1. In a social networking service (SNS), users generate raw data streams with emotions. Deep neural networks can capture the semantic emotion in the raw data streams and upload the raw data streams and the user's emotion label to the cloud using an SNS application.

Large volumes of data on social networks offer simple and efficient resources for performing emotional/sentiment analysis, and it has been shown that these resources can be leveraged for mental health monitoring [19][20]. As shown in Figure 1, the data flow of social media vlog data in social networks. Oftentimes, published data reflect not only the sentiment of the author, but also contain redundant information, such as events and activities [21]. In the application of depression recognition, semantic communication works to extract the true intention of the author, remove redundant information and retain only affective information related to depression. This simplifies the data, reduces the network load associated with communication, and facilitates the analysis of mental state. In recent years, transformer-based methods have played an irreplaceable role in depression detection research. However, most transformer-based methods do not take into account the spatio-temporal information in the data stream and the fusion of information between different modalities, which limits the capture of high-level semantic information in the data stream [22, 23, 24]. For example, although the correlation between key points in a sequence of facial landmarks is obvious, this spatial information is ignored in many studies compared to the temporal information. Our research is motivated by the further extraction of more effective depression-related spatio-temporal fusion features from facial and acoustic data streams.

In this work, we propose an end-to-end spatio-temporal squeezed transformer model to extract semantic features from social media vlog data for the task of identifying abnormal emotions such as depression and decrease the cost of communication transmission. The spatio-temporal squeezed transformer block (STST) is proposed for extracting temporal and spatial information from facial and vocal data. Specifically, the facial and vocal data in social media vlog data are first preprocessed before semantic feature extraction. Second, the spatio-temporal features in the data stream are extracted using the STST module embedded in value (V) of the cross-transformer encoder. Finally, a simple voting classifier is used to classify the extracted features. Figure 2(a) shows a schematic diagram of the framework of our proposed method for feature extraction. Our contributions and innovations to this field contained in this paper are as follows:

- A proposed method for automated detection of depression. Unlike other methodologies, ours extracts both temporal and spatial features from social media vlog data and fuses them in real time.
- Plug-and-play spatio-temporal squeeze fusion module, which is capable of capturing global information contained in the data while simultaneously focusing on local spatio-temporal features of the data.
- The proposed method exceeds the precision of the baseline method in the D-Vlog dataset, achieving state-of-the-art (SOTA) algorithmic performance.

The aim of this paper is to propose a spatio-temporal squeezed transformer approach to identify depression by extracting semantic information from facial and vocal streams in social media vlog data. Section 2 discusses the related work of the current study. In Section 3, the computational methods and models of our work is described. The experimental results and analysis are presented in Section 4. Section 5 gives the conclusion of this paper and future work.

2. Related Work

In this section, we provide a brief overview of previous work on SC in emotion recognition and Transformer and Self-Attention.

2.1. SC in Emotion Recognition

Due to the limitations of traditional communication technologies, some scholars have focused on edge computing to improve web service response (e.g., Telematics [25][26]), while others have recently focused on semantic computation to bypass bottlenecks. The concept of semantic communication was first proposed in 1952 [27], and acts as a complement to classical information theory. Bao et al. [28] proposed a reliable semantic communication model, using communication channels and information sources. Inspired by [28], Basu et al. [29] proved that parsimonious communication can be achieved by the semantic relationships between source symbols. To alleviate problems such as communication overload, Ning et al. [30] built a 5G home health monitoring system for Internet of Medical Things (IoMT) using the support of mobile edge computing.

Recently, deep learning (DL) has achieved great success in various fields, i.e., affective computing, computer vision, and Intelligent Transportation Systems (ITS) [31]. In the area of semantic communication, different types of semantic information impact model design. Therefore, we will provide a brief overview of different semantic communication models within text, speech, and image signals.

In text-based research, Guler et al. [32] turned the text-based semantic communication problem into a static Bayesian game and identified the Bayesian Nash equilibrium. This research revealed that semantic awareness in transmission can improve the communication performance, but it only considered the word level and ignored the sentence level. This drove Farsad et al. [33] to propose a joint source-channel coding architecture based on text, which extended sentences into semantic space. A recurrent neural network (RNN) was used as an encoder and decoder to decrease the word error rates (WER). Compared with traditional schemes, this DL method had a lower WER. Moreover, based on Transformer, a semantic communication model named DeepSC was proposed [34]. DeepSC can handle channel noise and semantic distortion, minimize the semantic errors, and recover semantic information.

Inspired by the previous text-based research, semantic communication systems for speech signals were developed. Weng et al. [6] proposed a speech-based system named DeepSC-S, where the channel coding and the speech coding were jointly designed. DeepSC-S utilized squeeze-and-excitation (SE) networks to recover transmitted speech signals and determined the transmission error at the semantic level instead of the bit or symbol levels.

In imaging applications, Bourtsoulatze et al. [35] first embedded CNN in a joint source and channel coding (JSCC) scheme, named deep JSCC, which directly mapped the input image to the channel. Kurka et al. [36] built on this work to develop DeepJSCC-f scheme for image, which utilized channel output feedback to improve reconstruction quality of image. Jankowski et al. [37] introduced pruning techniques into Deep JSCC to perform feature transmission with the goal to reduce redundancy in networked devices. Yang et al. [38] proposed a prototype of interference-free mental state assessment using facial video streams collected via 5G terminal to assess the mental state of users in real time.

2.2. Transformer and Self-Attention

The model structure of a Transformer was implemented by stacking multi-headed self-attention and feedforward multilayer perceptron (MLP) layers with residuals, which was first applied in the field of Natural Language Processing (NLP) [39]. The multi-headed attention mechanism captures the global information of the input sequence, while the residual structure combats the issue of disappearing gradients and degradation of the weight matrix. Due to its performance power, it has become increasingly popular for both sequential [40] and non-sequential tasks [41].

The birth of the Visual Transformer (VIT) [41] has revealed the potential of transformer-based image classification models. In recent years, it has replaced convolutional neural networks in various tasks in the field of computer vision tasks [42][43]. Research on multimodal fusion tasks [18, 44] has also been promoted (cross-transformer). Cross-transformer modeling is achieved by fusing single-modality information with each other by obtaining query (**Q**), key (**K**), and value (**V**) from different single-modality data. In order to capture the long-term contextual information in long-sequence audio/video of depression, a transformer encoder was applied to the depression detection task for the first time with excellent results [22]. Guo et al. [24] designed a two-branch transformer structure called Topic-Attentive Transformer-based (TOAT) for depression detection. In TOAT, features of single-modality samples are first extracted, and then a post-fusion strategy (simple feature stitching) is used to achieve multimodal informa-

tion fusion. However, this approach is not able to fuse features well, thus the stacking structure of the cross-transformer was used for efficient fusion [23]. Previous work has not taken the spatio-temporal features of the samples into account in the feature fusion strategy. In response, we seek to develop a method for fusing the spatio-temporal features of samples in real time to effectively identify depression.

3. Computational Methods and Models

In this section, we first define the research problem. Then, the structure of the transformer block is presented. Finally, a detailed description of the components of the proposed method is provided.

3.1. Problem Definition

Social media vlog data is used to accomplish the problem of affective computing based on semantic information to detect abnormal sentiment in this paper. User data contains data of two modalities with the same length of time, i.e., facial landmarks and vocal features data. $\mathbf{D} = (X_a, X_v)$ is used to represent the set of two single modalities of each user's data, where $X_a \in \mathbb{R}^{T\times N\times C}$ and $X_v \in \mathbb{R}^{T\times N\times C}$ denote facial landmarks and vocal features data respectively, and are represented as follows:

$$\begin{aligned} X_a &= \{X_{a_1}, X_{a_2}, \cdots X_{at}, \cdots X_{aT}\}, \\ X_v &= \{X_{v_1}, X_{v_2}, \cdots X_{vt}, \cdots X_{vT}\} \end{aligned} \tag{1}$$

where $X_{at} \in \mathbb{R}^{N\times C}$ and $X_{vt} \in \mathbb{R}^{N\times C}$ denote the facial landmarks and vocal features of a user at moment t, respectively. N and C denote the number and dimensionality of these values. Our task is to train an encoder to learn valid spatio-temporal representations of non-verbal features (acoustic and visual) extracted from social media vlog data streams to perceive depressive emotions. An overview of our approach is shown in Figure 2(a).

3.2. Dot-Product Self-Attention

The self-attentive mechanism is an important part of the transformer model. The self-attention mechanism computes the correlation between sequence tokens by mapping matrices and aggregates the information from different tokens. Formally, given queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$), attention is computed by the dot product, defined by the following equation:

$$\begin{aligned} \text{Att}\,(X) &= \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\ &= f_{\text{softmax}}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V} \end{aligned} \tag{2}$$

where $\mathbf{Q} = W^q X$, $\mathbf{K} = W^k X$, $\mathbf{V} = W^v X$, X denotes the input sequence for which attention needs to be calculated. d_k is the key dimensionality.

The multi-headed attention mechanism, proposed in [39], allows the model to focus on different distributed information, thus enhancing the expressive power of the model, as defined below:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}\,(\mathbf{H}_1, \ldots, \mathbf{H}_{n_h})\,\mathbf{W}^O \tag{3}$$

$$\mathbf{H}_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right) \tag{4}$$

where $\mathbf{W}_i^Q \in \mathbb{R}^{d_m\times d_k}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_m\times d_k}$, $\mathbf{W}_i^V \in \mathbb{R}^{d_m\times d_v}$, and $\mathbf{W}^O \in \mathbb{R}^{n_h d_v\times d_m}$ are the parameter matrix.

3.3. Method Overview

In this section, we will detail our method for the detection of emotional state, i.e., depression, for semantic information-based affective computing. The flow framework of our proposed method is illustrated in Figure 2(a). This includes the data preprocessing block, the token embedding block, the semantic feature extraction block (Stage 2), and the training the model. Semantic feature extraction blocks are stacked by K-layers of the same structure of the spatio-temporal squeezed transformer (STST) model, which is capable of focusing more on temporal and spatial information. The classifier of the model receives the feature representation of the sample from the semantic feature extraction block, which in turn predicts the label information of depression.

3.4. Data Preprocessing

Data preprocessing is necessary prior to feature extraction, as shown in Figure 2(b). Preprocessing includes data length and value normalization. The length of social media video data varies by user, in the D-Vlog dataset video length range from 23.62 and 3968.59 seconds. During data length normalization, the data are sliced according to length L, and data less than the length of L are complemented using polynomial interpolation. This operation is performed to increase the data size and prevent model overfitting. In this paper, the value of L is set to 300 frames. Values are normalization by using the maximum-minimum method, defined as follows:

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{5}$$

3.5. Token Generation

Before generating the token, we use a fully connected layer to project the input video stream data X into a new space X' (belonging to $\mathbb{R}^{T\times N\times C}$), which not only ensures uniformity of the data's dimensions in the feature space, but also enhances the representation capability of the model to some extent. The specific operations are as follows:

$$X' = \text{Reshape}\,(X) \tag{6}$$

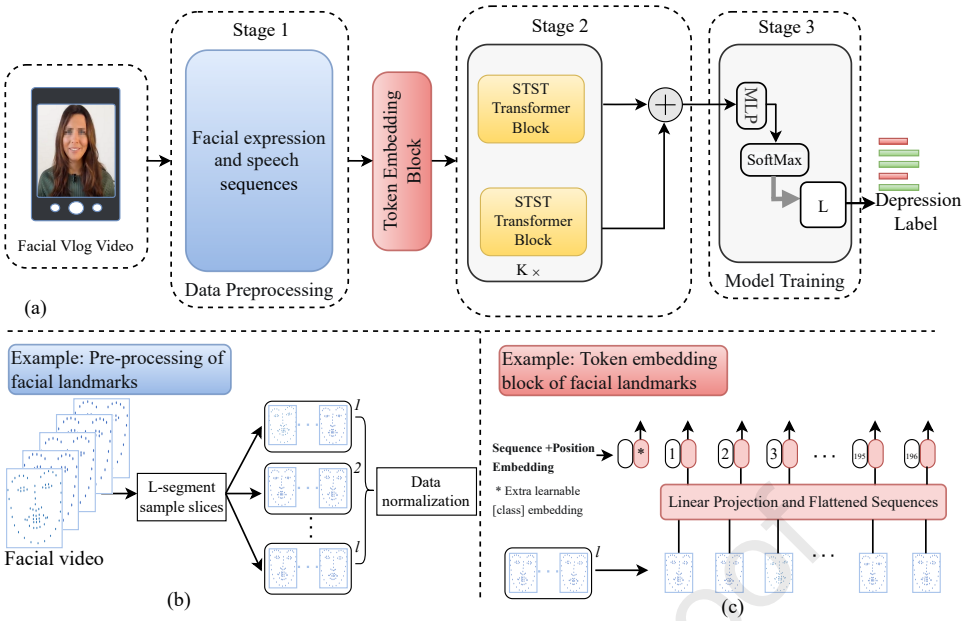


Fig. 2. (a) Flow chart for depression recognition. First, the non-verbal representations extracted from the video are pre-processed, then the spatio-temporal features in the data stream are extracted, and finally the depression classification is performed. (b) the flow of data preprocessing with facial landmarks as an example. (c) the specific process of token embedding with facial landmarks as an example.

where Reshape(·) denotes a two-layer fully connected operation.

Social media vlog data can be viewed as a stack of frame information in the temporal dimension; therefore, the user's sentiment changes depending on the temporal order. As shown in Figure 2(c), we append learnable location-encoded information to the input data sequence, represented as follows:

$$X = X + PE \tag{7}$$

where PE indicates the location-encoded information.

3.6. STST Encoder

Inspired by [45] and [46], we design the spatio-temporal squeezed transformer (STST), which has a cross-attentive mechanism with spatio-temporal information extraction capability, as shown in Figure 3. STST enables the model to extract both temporal and spatial information of social media vlog data by squeezing the spatio-temporal information into data blocks, which enables the query matrix (**Q**) to contain more comprehensive semantic information in the values (**V**) mapped in the keys (**K**).

In social media, temporal information can express the continuity of a user's emotions. We posit that the continuum of information is more dependent on the sentiment expressed by video frames that are temporally proximal than those that are more distant. The global temporal dependence can be achieved by adding location tokens to the data using an attention mechanism. In order to extract the relationship between video frames before and after the current moment, we emphasize temporal information highlighting the current frame. The importance of spatial information is emphasized using a spatial information extraction block, defined as follows:

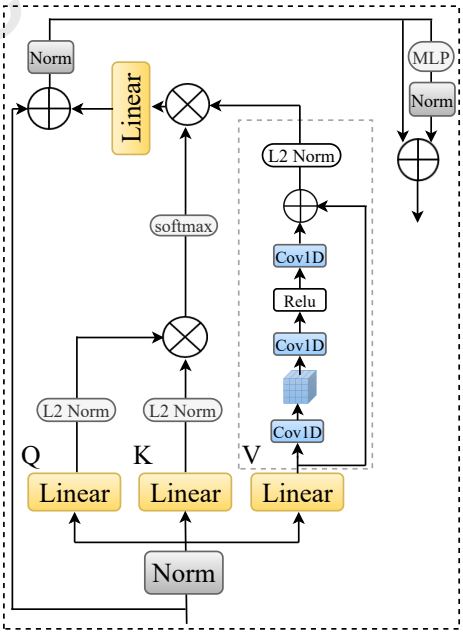


Fig. 3. Details of STST implementation

$$V_{ts} = \text{Space}(\text{Time}(V)) \tag{8}$$

where Time(·) is used to extract the temporal information from V and Space(·) is used to extract the spatial information from V. Specifically, they are implemented by one-dimensional convolution. The block of spatio-temporal information in V is put through a layer of activation functions that enable the model to learn and understand complex and non-linear data features,

defined as follows:

$$V_{Rule} = \text{Rule}(V_{ts}) \tag{9}$$

where Rule(·) [47] denotes the activation function. Then, a spatio-temporal information fusion layer is used to further fuse the spatio-temporal information extracted in the previous process, which can be written as Eq. 10.

$$V_{Fusion} = \text{Fusion}(V_{Rule}) \tag{10}$$

where Fusion(·) is the spatio-temporal feature fusion function, and we use a layer of one-dimensional convolution to implement it. To prevent the gradient from vanishing, the residual structure is introduced into the extraction process of spatio-temporal features, and the spatio-temporal feature squeeze block with residual terms can be written as Eq. 11.

$$V' = V_{Fusion} + V \tag{11}$$

The STST module is able to extract the spatio-temporal features of single-modality social data and fuse these features. Specifically, the attention of social media vlog data between single modalities is calculated using a cross-attention mechanism can be written as Eq. 12.

$$\begin{aligned} X'_a &= \text{Att}(X_a) = \text{MultiHead}(Q_v, K_v, V'_a), \\ X'_v &= \text{Att}(X_v) = \text{MultiHead}(Q_a, K_a, V'_v) \end{aligned} \tag{12}$$

where V'_a and V'_v are the facial landmarks and vocal features data, with spatio-temporal information calculated by Eq. 3. X_v and X_a are calculated by linear transformation to obtain Q_v, K_v, and V'_a.

$$\begin{aligned} Q_v &= \phi(X_v), \\ K_v &= \phi(X_v), \\ V'_a &= \text{Fusion}(\phi(X_a)) + V_a \end{aligned} \tag{13}$$

Similarly, Q_a, K_a, and V'_v can also be calculated.

Then, the same calculation is performed as in the transformer encoder, i.e., an MLP network with residual structure.

$$\begin{aligned} X_a^{out} &= \text{MLP}(X'_a) + X'_a, \\ X_v^{out} &= \text{MLP}(X'_v) + X'_v \end{aligned} \tag{14}$$

Eq. 12, 13, and 14 are combined to obtain Eq. 15:

$$\begin{aligned} X_a^{out} &= \text{STST}\left(X_a, X_v\right), \\ X_v^{out} &= \text{STST}\left(X_v, X_a\right) \end{aligned} \tag{15}$$

where STST(·) is used to extract the spatio-temporal fusion features of the single-modality data.

3.7. Model Training

To verify the effectiveness of the semantic features extracted by the proposed model, we designed a depression classifier consisting of an MLP layer and a Softmax layer. The MLP network layer consists of two linear fully connected layers. The depression classifier allows for the predictive labels $\hat{y}$, defined as follows:

$$\hat{y} = \text{MLP}\left(\text{Concat}(X_a^{out}, X_v^{out})\right) \tag{16}$$

The duration of video data varies from user to user, so each user's social media vlog data is divided into a different number of slices in the data preprocessing stage. Assuming that the number of slices into which a user's social media vlog data is divided is *Num*, feeding *Num* slices into the depression classifier will result in *Num* predictive labels that are not necessarily identical. To determine the unique affective state of the user, we take the category with the most predicted categories among the *Num* predictive labels as the affective state of the current user. Figure 4 shows the facial expressions (e.g., happy, angry, etc.) shown at different moments in the vlog videos of depressed users. Different facial expressions express different emotions, which may reduce the classification accuracy. Therefore, a depression classifier with a voting mechanism that can account for the ratios of emotions can alleviate this problem.

In this paper, the cross-entropy loss function is used as the loss function for the model calculation and is defined as follows:

$$L = \frac{1}{N}\sum_{j} -\left[y_j \cdot \log\left(p_j\right) + \left(1 - y_j\right) \cdot \log\left(1 - p_j\right)\right] \tag{17}$$

where y_j denotes the label of social media vlog data j, the positive class is 1 (depression), and the negative class is 0 (non-depression). p_j denotes the probability that social media vlog data j is predicted to be in the positive class. Algorithm 1 reveals the training scheme of our proposed method.

Fig. 4. Different expressions in vlog data of depressed patients

4. Experimental Results and Analysis

In this section, we present the data set and experimental details and validate the model from several perspectives.

Algorithm 1 Training Scheme of Our Proposed Method

Input: Multimodal depression dataset $\mathcal{D} = \{(\mathbf{D}, \boldsymbol{y})\}$, where $\mathbf{D} = (X_a, X_v)$.
Output: Prediction $\hat{\boldsymbol{y}}$.

for $i = 1$ to Epoches **do**
  for $k = 1$ to K **do**
    Capture facial and vocal features:
    $X_a^{out} = \mathrm{MLP}(X_a') + X_a'$, where $X_a' = \mathrm{MultiHead}(Q_v, K_v, \mathrm{Fusion}(\phi(X_a)) + \phi(X_a))$;
    $X_v^{out} = \mathrm{MLP}(X_v') + X_v'$, where $X_v' = \mathrm{MultiHead}(Q_a, K_a, \mathrm{Fusion}(\phi(X_v)) + \phi(X_v))$.
  end for
  Compute the final label $\hat{y}$: $\hat{y} = \mathrm{MLP}\left(\mathrm{Concat}(X_a^{out}, X_v^{out})\right)$.
  Compute loss L.
  Backward L and update parameters.
end for

4.1. Datasets and Experiment Setting

4.1.1. Datasets

We evaluate our proposed method on Depression Vlog (D-Vlog) depression dataset [23]. This dataset contains 916 (i.e., approximately 160 hours) of vlog video clips collected from 816 users on YouTube, including 555 depression data and 406 non-depression data. Yoon et al. [23] analyzed and processed the vlog videos, which were collected on YouTube between 2020.1.1 and 2021.1.31, based on specific keywords(i.e., depression and non-depression keywords), and finally obtained the vlog dataset with labels. The proposed D-Vlog dataset has been validated by DAIC-WOZ [48] (a clinically labeled dataset) as useful for depression research. The D-Vlog dataset is divided into three parts: train, validation and test sets, as shown in Table 1 and to avoid gender bias, the sections are divided according to the proportion of depression and non-depression by gender. Two types of data are included in the dataset, facial landmarks and acoustic features; they are extracted from the raw vlog data using the dlib toolkit [49] and opensmile [50] toolkit respectively, which can effectively protect the user's privacy.

Table 1
The number of sample divisions in the D-Vlog dataset [23].

Gender	Train	Validation	Test
Male	216	40	66
Female	431	62	146

4.1.2. Implementation Details

All experiments are conducted using the Pytorch [51] framework, and model training is performed on two NVIDIA Tesla V100-PCIE GPUs each with 32GB memory. We set the value K of the stacked blocks in stage 2 of Figure 2(a) to 2 and the number of cross-attention heads to 4. To avoid the impact of invalid data on the model performance, data with all-zero values in the D-Vlog dataset are removed (e.g., for vlog data with null values for all facial landmarks).

Token embedding blocks use learnable token encoding strategies. The kernel size of the 1x1 convolution in the STST module is 3. The batch size and epochs of our model are 64 and 600, respectively. This model uses an SGD [52] optimizer with a momentum of 0.5, a weight decay of 0.001 and a learning rate with a cosine learning decay [53] with an initial value of 0.001. To prevent overfitting, we use a dropout strategy with a dropout [54] rate of 0.5 in the linear layer, the position encoding, and the transformer encoder. Precision, recall, f1 score, and accuracy are used as evaluation metrics to assess model performance.

4.2. Parameters of the Model

We perform selection experiments on the number of stacks K of the STST module and the token generation method of the samples to ensure the effectiveness of our method.

Table 2
The model STST stacking layer on the value of K for activation selection.

K	Acc(%)	Pre(%)	Rec(%)	F1(%)
1	69.19	**72.56**	73.21	72.88
2	**70.70**	72.50	77.67	**75.00**
3	64.64	67.50	72.32	69.82
4	67.67	70.00	75.00	72.41
5	65.15	65.92	**79.46**	72.06
6	65.65	70.00	68.75	69.36

The choice of K values in our method is important because too small a value of K will lead to underfitting the model and it will fail to learn the feature expression of the sample; too large a value will lead to overfitting the model and affect model performance on the validation data. In order to investigate the effect of the magnitude of the K value on the model's performance, we conducted experiments on the choice of K value. As shown in Table 2, the model achieves the highest accuracy when K = 2. It is worth noting that

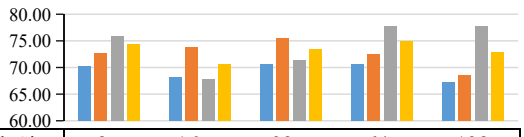
when K = 1, the accuracy rate is 69.19%, second only to the accuracy rate when K = 2, which indicates that the model is underfitting. The accuracy of the model decreases severely when K = 3 compared to 70.70%, which indicates that the model is overfitting. The F1 value (F1) is the summed average of precision (Pre) and recall (Rec). If only Pre or only Rec is considered, neither can be used as an indicator to evaluate a good or bad model. In Table 2, when K = 2, the F1 is at its maximum (75.00%), which indicates the model performance is most stable.

Table 3
The impact of different location codes on the performance of this model.

Position Encoder	Acc(%)	Pre(%)	Rec(%)	F1(%)
sinusoidal	61.61	63.63	75.00	68.85
learned	**70.70**	**72.50**	**77.67**	**75.00**

The token generation module appends a set of values to the samples, allowing the model to learn global inter-frame position information. In order to select the token generation method that is more suitable for our approach, we compared 'sinusoidal' and 'learned' token generation approaches. As shown in Table 3, the evaluation metrics from 'sinusoidal' token generation are lower than those of the 'learned'. There is a significant difference in the magnitude of accuracy, which may occur because the 'learned' token generation approach is more suitable for our model to extract the sequence information of facial landmarks and vocal features in social media vlog data.

At a learning rate of 0.001 and a stacking number K of 2, the effect of batch size on model prediction results. Each evaluation index of the model is more stable for batch size of 64 than for other batch size, as shown in Figure 5.



Batch Size	8	16	32	64	128
Acc(%)	70.20	68.18	70.70	70.70	67.17
Pre(%)	72.64	73.78	75.47	72.50	68.50
Rec(%)	75.89	67.85	71.42	77.67	77.67
F1(%)	74.32	70.69	73.39	75.00	72.80

Fig. 5. The effect of batch size on model prediction results at a learning rate of 0.001 and a stacking number K of 2.

4.3. Ablation

We conduct ablation experiments to demonstrate the effectiveness of our method. All experiments are performed on a batch size 64 with a learning rate 0.001. All the ablation experiments are validated on the results of the test set.

Validity of spatio-temporal feature extraction. We perform ablation experiments to provide validity of each of the model's components. These are elimination of: the temporal feature extraction block (STSTs^-), the spatial feature extraction block (STSTt^-), and spatio-temporal feature extraction block (STST), to explore the role assumed by each component in our method. As shown in Table 4, the accuracy of using only STSTt^- is lower than that of using only STSTs^-, which indicates that temporal features are better than spatial features in distinguishing depressed from non-depressed populations, which can be considered dependent on time-varying facial landmarks and vocal features compared to spatial features.

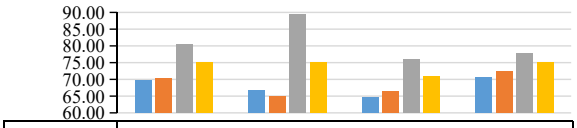
Table 4
Ablation studies with different module performance results.

Time	Space	Acc(%)	Pre(%)	Rec(%)	F1(%)
	✓	63.63	70.00	62.50	66.03
✓		67.67	69.67	75.89	72.64
✓	✓	**70.70**	**72.50**	**77.67**	**75.00**

As shown in Table 4, the accuracy of using STST is higher than that of using STSTs^- alone or the STSTt^- alone. This indicates that although the spatial information of the samples is not as important as the temporal information, the mutual integration of temporal and spatial features (i.e., the simultaneous extraction of spatio-temporal features in the samples), improves model performance.

Validity of models with different structures. Fang et al. [55] proposed a spatio-temporal feature extraction block (STJA) for use in traffic prediction tasks. They applied STJA to the query (Q) and key (K) in the transformer encoder to extract spatio-temporal features in traffic data. The design sought to extract the spatio-temporal information in Q and K and map it to the values in V more efficiently. Learning from this idea, we design spatio-temporal feature extraction blocks (STST-QKst) focusing on Q and K to demonstrate the effectiveness of our model. Additionally, for the order of spatio-temporal feature extraction, experiments are designed to first extract spatial features and then temporal features (STSTst). For the connection of spatio-temporal features, experiments are designed to extract temporal and spatial features of samples separately and fuse them in a simple summarization manner (STSTadd).

As shown in Figure 6, comparing the accuracy of STST-QKst reveals that the model that embeds spatio-temporal attention into Q and K and extracts spatial features first performs better than the model that extracted temporal features first. This is contrary to



Model	STST-QKst	STSTst	STSTadd	STST
Acc(%)	69.60	66.66	64.64	70.70
Pre(%)	70.31	64.93	66.40	72.50
Rec(%)	80.35	89.28	75.89	77.67
F1(%)	75.00	75.18	70.83	75.00

Fig. 6. Ablation studies with different module performance results.

the conclusion that temporal information is extracted first, and then spatial features are extracted in our method. The possible reason for this phenomenon in these two sets of experimental results is that during the computation of the attention mechanism in the transformer encoder, a matrix transpose is performed before the matrix V is involved in the computation. It is worth noting that the experimental accuracy of STST-QKst closely approximates our proposed STST model, which indicates that STST-QKst is also a better model. However, from the perspective of computational power, our proposed method has less computational power compared to STST-QKst. The above conclusions can be confirmed further in the experimental STSTst, whose accuracies are close to and significantly lower than the experimental STST. The accuracy of STSTadd is lower than our method, probably because a simple summation will cancel out some weighted features, which is not conducive to the effective learning of the model for the sample features. To highlight this, normalized confusion matrices of models STST and STST-KQst are shown in Figure 7.

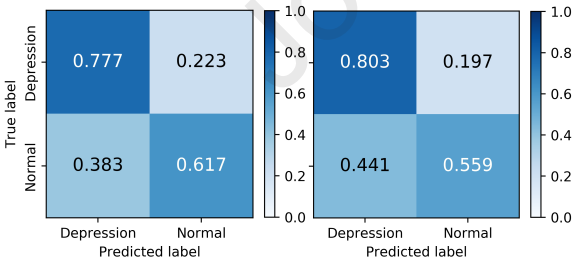


Fig. 7. Normalized confusion matrices of models STST and STST-KQst are shown in the right subplot and left subplot, respectively. Each row of the confusion matrices represents the true label and each column represents the predicted label. The element (i, j) is the percentage of samples in class i that is classified as class j.

4.4. Comparison with Previous Methods

To validate the performance of our approach, we compared seven baseline models, including traditional machine learning and deep learning. The traditional machine learning methods are Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and k-Nearest Neighbor-based Fusion (KNN-Fusion [56]). Deep learning models include Bi-directional LSTM (BLSTM), Tensor Fusion Network (TFN [57]), and Depression Detector[23]. KNN-Fusion, proposed by Pampouchidou et al., is a machine learning method for depression detection using decision-level fusion methods to fuse audio and video. BLSTM has been shown to efficiently extract features of time series to detect depression [58]. Specifically, two single modalities data are input to two BLSTMs to extract effective features, and a fully connected layer with softmax is used as a classifier. TFN is a network structure in multimodal sentiment analysis. The Depression Detector is a multimodal fusion depression detection network based on a cross-transformer encoder.

To facilitate comparison with previous methods, we use weighted average precision, recall, and F1 score as the evaluation metrics of the models. As shown in Table 5, the performance of the deep learning models is better than that of the traditional machine learning methods. This may be due to the fact that traditional machine learning requires the artificial extraction of some manual features, which do not represent the original samples well. Compared with other methods, our method obtains excellent performance, i.e., precision of 72.50%, recall of 77.67%, and F1 score of 75.00%. This is due to the ability of the STST model, which simultaneously extracts spatio-temporal features of the samples and use a classifier with a voting mechanism that can synthesize the total tone of the user's emotions throughout the video time. The experimental results show that our approach is effective in using D-vlog data to distinguish between depressed and non-depressed populations.

Table 5
Performance comparisons between baseline models and the proposed model.

Model	Pre(%)	Rec(%)	F1(%)
LR	54.86	54.72	54.78
SVM	53.10	55.19	52.97
RF	57.69	58.49	57.84
KNN-Fusion	57.86	59.43	54.25
BLSTM	60.81	61.79	59.70
TFN	61.39	62.26	61.00
Depression Detector	65.40	65.57	63.50
STST(Ours)	72.50	77.67	75.00

5. Conclusions and Future Work

In this paper, we propose a spatial-temporal squeeze transformer framework for extracting spatial-temporal features from social media vlog data in semantic communication to recognition depression. We focus on the temporal and spatial features of user data and fuse

these features in real time to facilitate the extraction of features that are more easily distinguishable between depressed and non-depressed people. Experimental results show that the performance of our method is significantly better than other methods across evaluation metrics. Moreover, the method of using multimodal fusion to extract semantic information offers a framework for future research.

There are some limitations to the current study. The expression of emotion is closely related to individual habits and has a high degree of individual correlation. In order to address this shortcoming, it may be possible to utilize contrast learning. That is, while using a supervised model to extract features used for semantic communication, an unsupervised contrast learning model is used to learn the emotional expression habits of individuals. Further, multimodal fusion can improve the effectiveness of the extracted features and facilitate depression recognition. Therefore, in future work, more single-modality fusion strategies will be used in semantic communication tasks to obtain more effective semantic information.

References

[1] R. Dangi, P. Lalwani, G. Choudhary, I. You, G. Pau, Study and investigation on 5g technology: A systematic review, Sensors 22 (1) (2021) 26.

[2] Z. Ning, P. Dong, X. Kong, F. Xia, A cooperative partial computation offloading scheme for mobile edge computing enabled internet of things, IEEE Internet of Things Journal 6 (3) (2018) 4804–4814.

[3] Y. Zhang, C. Jiang, B. Yue, J. Wan, M. Guizani, Information fusion for edge intelligence: A survey, Information Fusion 81 (2022) 171–186.

[4] W. Weaver, Recent contributions to the mathematical theory of communication, ETC: a review of general semantics (1953) 261–281.

[5] S. Jiang, Y. Liu, Y. Zhang, P. Luo, K. Cao, J. Xiong, H. Zhao, J. Wei, Reliable semantic communication system enabled by knowledge graph, Entropy 24 (6) (2022) 846.

[6] Z. Weng, Z. Qin, G. Y. Li, Semantic communications for speech signals, in: ICC 2021-IEEE International Conference on Communications, IEEE, 2021, pp. 1–6.

[7] Q. Zhou, R. Li, Z. Zhao, C. Peng, H. Zhang, Semantic communication with adaptive universal transformer, IEEE Wireless Communications Letters 11 (3) (2021) 453–457.

[8] X. Chen, M. D. Sykora, T. W. Jackson, S. Elayan, What about mood swings: Identifying depression on twitter with temporal measures of emotions, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 1653–1660.

[9] Z. A. E. Sarhan, H. A. El Shinnawy, M. E. Eltawil, Y. Elnawawy, W. Rashad, M. S. Mohammed, Global functioning and suicide risk in patients with depression and comorbid borderline personality disorder, Neurology, Psychiatry and Brain Research 31 (2019) 37–42.

[10] G. S. Malhi, J. J. Mann, Depression, The Lancet 392 (10161) (2018) 2299–2312.

[11] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, D. S. Geralts, Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology, Journal of neurolinguistics 20 (1) (2007) 50–64.

[12] W. Wang, X. Yu, B. Fang, D.-Y. Zhao, Y. Chen, W. Wei, J. Chen, Cross-modality LGE-CMR segmentation using image-to-image translation based data augmentation, IEEE/ACM Transactions on Computational Biology and Bioinformatics 2022.

[13] J. Chen, S. Sun, L.-b. Zhang, B. Yang, W. Wang, Compressed sensing framework for heart sound acquisition in internet of medical things, IEEE Transactions on Industrial Informatics 18 (3) (2022) 2000–2009.

[14] A. Jan, H. Meng, Y. F. B. A. Gaus, F. Zhang, Artificial intelligent system for automatic depression level analysis through visual and vocal expressions, IEEE Transactions on Cognitive and Developmental Systems 10 (3) (2017) 668–680.

[15] L. Wen, X. Li, G. Guo, Y. Zhu, Automated depression diagnosis based on facial dynamic analysis and sparse coding, IEEE Transactions on Information Forensics and Security 10 (7) (2015) 1432–1441.

[16] B. Stasak, D. Joachim, J. Epps, Breaking age barriers with automatic voice-based depression detection, IEEE Pervasive Computing 21 (2) (2022) 10–19.

[17] L. He, M. Niu, P. Tiwari, P. Marttinen, R. Su, J. Jiang, C. Guo, H. Wang, S. Ding, Z. Wang, et al., Deep learning for depression recognition with audiovisual cues: A review, Information Fusion 80 (2022) 56–86.

[18] H. Sun, Y.-W. Chen, L. Lin, Tensorformer: A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection, IEEE Transactions on Affective Computing, 2022 (2022).

[19] S. Gupta, L. Goel, A. Singh, A. Prasad, M. A. Ullah, Psychological analysis for depression detection from social networking sites, Computational Intelligence and Neuroscience 2022 (2022).

[20] U. Ahmed, J. C.-W. Lin, G. Srivastava, Social media multi-aspect detection by using unsupervised deep active attention, IEEE Transactions on Computational Social Systems (2022) (2022).

[21] S. Ghosh, A. Ekbal, P. Bhattacharyya, What does your bio say? inferring twitter users' depression status from multimodal profile information using deep learning, IEEE Transactions on Computational Social Systems (2021) (2021).

[22] H. Sun, J. Liu, S. Chai, Z. Qiu, L. Lin, X. Huang, Y. Chen, Multi-modal adaptive fusion transformer network for the estimation of depression level, Sensors 21 (14) (2021) 4764.

[23] J. Yoon, C. Kang, S. Kim, J. Han, D-vlog: Multimodal vlog dataset for depression detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

[24] Y. Guo, C. Zhu, S. Hao, R. Hong, A topic-attentive transformer-based model for multimodal depression detection, arXiv preprint arXiv:2206.13256, 2022 (2022).

[25] Z. Ning, K. Zhang, X. Wang, L. Guo, X. Hu, J. Huang, B. Hu, R. Y. Kwok, Intelligent edge computing in internet of vehicles: a joint computation offloading and caching solution, IEEE Transactions on Intelligent Transportation Systems 22 (4) (2020) 2212–2225.

[26] Z. Ning, J. Huang, X. Wang, J. J. Rodrigues, L. Guo, Mobile edge computing-enabled internet of vehicles: Toward energy-efficient scheduling, IEEE Network 33 (5) (2019) 198–205.

[27] R. Carnap, Y. Bar-Hillel, et al., An outline of a theory of semantic information, Research Laboratory of Electronics, Massachusetts Institute of Technology (1952) (1952).

[28] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, J. A. Hendler, Towards a theory of semantic communication, in: 2011 IEEE Network Science Workshop, IEEE, 2011, pp. 110–117.

[29] P. Basu, J. Bao, M. Dean, J. Hendler, Preserving quality of information by using semantic relationships, Pervasive and Mobile Computing 11 (2014) 188–202.

[30] Z. Ning, P. Dong, X. Wang, X. Hu, L. Guo, B. Hu, Y. Guo, T. Qiu, R. Y. Kwok, Mobile edge computing enabled 5g health monitoring for internet of medical things: A decentralized game theoretic approach, IEEE Journal on Selected Areas in Communications 39 (2) (2020) 463–478.

[31] Z. Ning, K. Zhang, X. Wang, M. S. Obaidat, L. Guo, X. Hu, B. Hu, Y. Guo, B. Sadoun, R. Y. Kwok, Joint computing and caching in 5g-envisioned internet of vehicles: A deep reinforcement learning-based traffic control system, IEEE Trans-

actions on Intelligent Transportation Systems 22 (8) (2020) 5201–5212.

[32] B. Güler, A. Yener, A. Swami, The semantic communication game, IEEE Transactions on Cognitive Communications and Networking 4 (4) (2018) 787–802.

[33] N. Farsad, M. Rao, A. Goldsmith, Deep learning for joint source-channel coding of text, in: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2018, pp. 2326–2330.

[34] H. Xie, Z. Qin, G. Y. Li, B.-H. Juang, Deep learning enabled semantic communication systems, IEEE Transactions on Signal Processing 69 (2021) 2663–2675.

[35] E. Bourtsoulatze, D. B. Kurka, D. Gündüz, Deep joint source-channel coding for wireless image transmission, IEEE Transactions on Cognitive Communications and Networking 5 (3) (2019) 567–579.

[36] D. B. Kurka, D. Gündüz, Deepjscc-f: Deep joint source-channel coding of images with feedback, IEEE Journal on Selected Areas in Information Theory 1 (1) (2020) 178–193.

[37] M. Jankowski, D. Gündüz, K. Mikolajczyk, Joint device-edge inference over wireless links with pruning, in: 2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), IEEE, 2020, pp. 1–5.

[38] M. Yang, Y. Ma, Z. Liu, H. Cai, X. Hu, B. Hu, Undisturbed mental state assessment in the 5g era: a case study of depression detection based on facial expressions, IEEE Wireless Communications 28 (3) (2021) 46–53.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017) (2017).

[40] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap, A. Sriram, V. Liptchinsky, R. Collobert, End-to-end asr: from supervised to semi-supervised learning with modern architectures, arXiv preprint arXiv:1911.08460, 2019 (2019).

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2020.

[42] J. Yang, X. Dong, L. Liu, C. Zhang, J. Shen, D. Yu, Recurring the transformer for video action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14063–14073.

[43] C. Wang, Z. Wang, Progressive multi-scale vision transformer for facial action unit detection, Frontiers in Neurorobotics 15 (2021) (2021).

[44] A.-M. Bucur, A. Cosma, P. Rosso, L. P. Dinu, It's just a matter of time: Detecting depression with time-enriched multimodal transformers, arXiv preprint arXiv:2301.05453, 2023 (2023).

[45] C. Doersch, A. Gupta, A. Zisserman, Crosstransformers: spatially-aware few-shot transfer, Advances in Neural Information Processing Systems 33 (2020) 21981–21993.

[46] B. Li, P. Xiong, C. Han, T. Guo, Shrinking temporal attention in transformers for video action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

[47] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[48] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al., The distress analysis interview corpus of human and computer interviews, Tech. rep., University of Southern California Los Angeles (2014).

[49] D. E. King, Dlib-ml: A machine learning toolkit, The Journal of Machine Learning Research 10 (2009) 1755–1758.

[50] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE transactions on affective computing 7 (2) (2015) 190–202.

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019) (2019).

[52] L. Bottou, Stochastic gradient descent tricks, in: Neural networks: Tricks of the trade, Springer, 2012, pp. 421–436.

[53] I. Loshchilov, F. Hutter, Sgdr: Stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983, 2016 (2016).

[54] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research 15 (1) (2014) 1929–1958.

[55] Y. Fang, F. Zhao, Y. Qin, H. Luo, C. Wang, Learning all dynamics: Traffic forecasting via locality-aware spatio-temporal joint transformer, IEEE Transactions on Intelligent Transportation Systems (2022) (2022).

[56] A. Pampouchidou, O. Simantiraki, C.-M. Vazakopoulou, C. Chatzaki, M. Pediaditis, A. Maridaki, K. Marias, P. Simos, F. Yang, F. Meriaudeau, et al., Facial geometry and speech analysis for depression detection, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017, pp. 1433–1436.

[57] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1103–1114.

[58] S. Yin, C. Liang, H. Ding, S. Wang, A multi-modal hierarchical recurrent neural network for depression detection, in: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 65–71.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: