DepMGNN: Matrixial Graph Neural Network for Video-based Automatic Depression Assessment

Zijian Wu[1, 2, 3], **Leijing Zhou**[4, 1], **Shuanglin Li**[5], **Changzeng Fu**[6], **Jun Lu**[3]
Jing Han[3], **Yi Zhang**[3], **Zhuang Zhao**[3], **Siyang Song**[2*]
[1]Affect AI, Anhui, China
[2]University of Exeter, UK
[3]Nanjing University of Science and Technology, China
[4]Zhejiang University, China
[5]University of Newcastle-upon-Tyne, UK
[6]Osaka University, Japan

Abstract

Depression can be reflected by long-term human spatio-temporal facial behaviours. While human face videos recorded in real-world usually have long and variable lengths, existing video-based depression assessment approaches frequently re-sample/down-sample such videos to short and equal-length videos, or split each video into several equal-length segments, where segment-level spatio-temporal facial behaviours are suppressed as a vector-style representations for RNN-based long-term (video-level) modelling. Both strategies lead to crucial information loss and distortion. In this paper, we propose a novel graph-style data structure called Matrixial Graph and an effective Matrixial Graph Neural Network (MGNN) for face video-based depression assessment, which can directly and end-to-end model long-term depression-specific spatio-temporal facial cues from variable-length videos without resampling/splitting videos or suppressing video segments to vectors. Importantly, the nodes in our matrixial graph are capable of including matrices of different shapes, and thus nodes of a matrix graph can directly represent all frame-level 2D facial feature maps (or images themselves) of an entire video regardless of its length. Then, our MGNN is the first GNN that can jointly process matrixial graphs containing varying numbers of nodes, which further learns matrix-style edge features, thereby facilitating to explicit model video-level multi-scale spatio-temporal facial behaviours among matrixial graph nodes for depression assessment. Experiments show that the explicit spatio-temporal modeling on 2D facial feature maps, facilitated by our matrixial graph/MGNN, provided significant benefits, leading our approach to achieve new state-of-the-art performances on AVEC2013 and AVEC2014 datasets with large advantages.

Code — https://github.com/AffectAI/MGNN

1 Introduction

Depression is a common but serious mental health problem that negatively impacts more than 300 millions individuals' daily life all over the world (James et al. 2018). Traditional depression assessments usually rely on questionnaires and interviews conducted by psychologists, which are subjective, expensive and time-consuming (Song et al. 2022). Since there are well-documented differences in spatio-temporal facial behaviours expressed by depressed and non-depressed individuals (e.g., reduced positive facial displays (Clark and Watson 1991) and lack of facial expressions (Ellgring 2007)), machine learning (ML) models have been frequently developed for automatic depression assessment (ADA) from human face videos (Valstar et al. 2014; Ringeval et al. 2019, 2017; He et al. 2022).

Early video-based ADA approaches typically rely on hand-crafted facial features (Meng et al. 2013; Pampouchidou et al. 2016), which fail to include depression-specific facial cues that are not considered in their manually defined feature extraction processes. Consequently, recent advances in deep learning (DL) have been widely applied. Given a face video (usually less than one hour (Valstar et al. 2013; Ringeval et al. 2019)), a typical DL-based solution starts with independently learning a depression-specific static feature from each face frame, and then conducts either feature-level or decision-level fusion of all frames to make the final video-level depression prediction (He, Chan, and Wang 2021; Shang et al. 2021). The main issue of such approaches is that crucial temporal dependencies among frames are not considered for their depression assessment.

As a result, some approaches employed Long-short-term-memory (LSTM) (Ray et al. 2019; Fang et al. 2023; Al Hanai, Ghassemi, and Glass 2018) to model depression-specific temporal dependencies from frame-level facial features, despite that crucial depression-specific spatio-temporal cues might be disregarded in their static frame-level feature extraction process. To avoid this, spatio-temporal Convolution Neural Networks (CNNs) have been widely applied to directly learn depression-specific spatio-temporal cues from the given video/video segment (Song et al. 2024; Al Jazaery and Guo 2018; de Melo, Granger, and Hadid 2020; Xu et al. 2024; Niu et al. 2022b). Since regular DL models can only process inputs with the same and fixed size (e.g., videos of the same length) while real-world face videos usually have variable lengths, these approaches frequently split every target video into multiple equal-length segments and learn a vector-style feature from each segment, which are finally combined to predict video-

*Corresponding author (ss2796@cam.ac.uk).

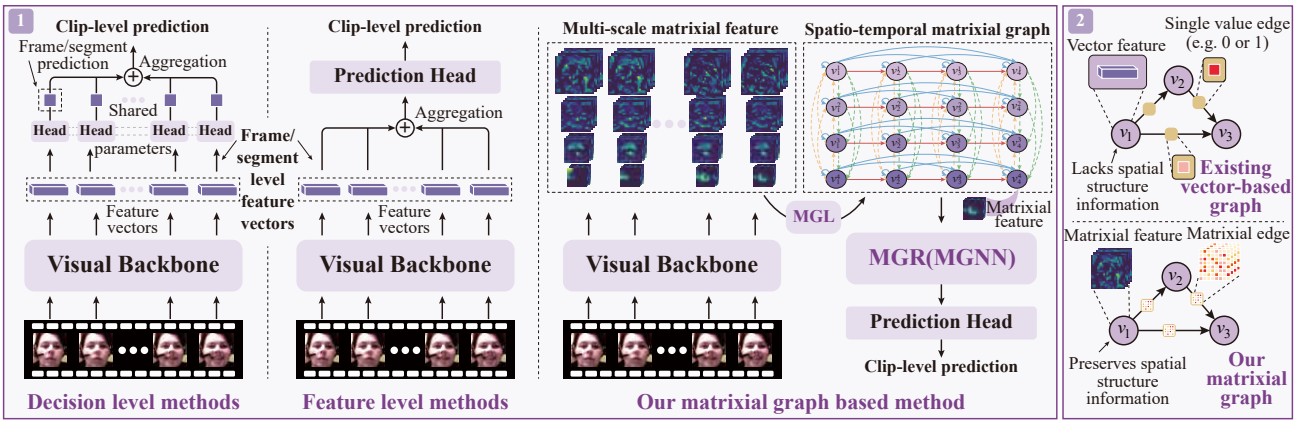

Figure 1: **(1) Existing video-based ADA approaches vs. our approach. Left/Middle**: Existing ADA approaches represent each frame/segment-level behaviour as a feature vector, and then aggregate their predictions (decision-level fusion) or features (feature-level fusion) to make a clip-level prediction, which fail to consider depression-specific spatio-temporal facial cues among frames/segments during their frame/segment-level feature extraction. **Right**: Our approach directly includes multi-scale 2D facial feature maps of each frame as a set of nodes in our novel matrixial graph as the clip-level representation, facilitating clip-level multi-scale spatio-temporal depression-specific feature extraction. **(2) Existing vector-based graphs vs. our matrixial graph.** In existing graphs, each node can not only include a vector (or a single value), while each node in our matrixial graph can include a a set of matrices and thus can better represent spatial cues.

level depression status. However, such approaches not only frequently remove some extra frames as they are not enough to form a individual segment meeting the pre-defined length, i.e., they fail to utilize all frames provided in the clip for the depression assessment (**Problem 1**), but also only able to model video-level temporal dependencies among segment-level predictions or features, which may ignore some crucial long-term depression-specific spatio-temporal cues in segment-level feature extraction (**Problem 2**). Although several ADA solutions (de Melo, Granger, and Lopez 2021; Pan et al. 2024) can model video-level spatio-temporal cues from the entire video, they have to largely downsample each video as a short facial image sequence, which not only distort video-level temporal information but also lose informative short-term facial dynamics (**Problem 3**).

In this paper, we propose a novel graph-style data called **Matrixial Graph** that can directly represent matrix-style representations of all frames (i.e., images themselves or their 2D feature maps) in an arbitrary-length face video, where each frame-level matrix is directly included in a matrixial graph node. This way, videos of varying frame numbers can be directly represented by matrixial graphs containing different numbers of nodes, without requiring to re-sample videos, remove frames or suppress frames to feature vectors. Accordingly, a novel **Matrixial Graph Neural Network (MGNN)** is also proposed, which assigns task-specific matrix-style spatial and temporal edges to connect matrixial graph nodes, facilitating explicit modelling of multi-scale spatio-temporal facial behaviours among frame-level matrixial representations included in matrix graph nodes for depression assessment. Fig. 1 compares our approach with previous solutions. The main contributions and novelties of this paper are summarised as follows:

- We propose a novel and effective matrixial graph-based ADA approach that can directly model video-level depression-specific spatio-temporal facial cues from all original frames or their 2D facial feature maps of an arbitrary-length video without discarding any of them or suppressing them to vectors.
- We propose a novel graph-style data structure: matrix graph, whose nodes can represent matrices of different shapes, and thus allows to directly include video frames (or their multi-scale 2D feature maps) without suppressing them to vectors as standard graphs.
- We propose the first GNN (called MGNN) that can directly process matrixial graphs containing variable numbers of nodes whose features are matrices, which differs from existing GNNs (Dwivedi et al. 2023) that can only process graphs whose node features are vectors/single values, our MGNN also innovatively learn matrix-style edge features to comprehensively control message passing among matrixial nodes during reasoning.

2 Related Work

Video-based ADA: Human facial behaviours are informative depression biomarkers (Clark and Watson 1991; Ellgring 2007). Given a face video, some studies (He, Chan, and Wang 2021; Shang et al. 2021) deep learn depression-specific cues from each of its face frames. They typically start with extracting static features from each facial frame/patch, and then aggregate all frame-level features/predictions to obtain the video-level depression score (e.g., averaging all frame-level predictions (Zhou et al. 2018)). Since such approaches ignore crucial temporal dependencies among facial frames, Uddin et al. (Uddin, Joolee, and Lee 2020) further introduced LSTM to model them among

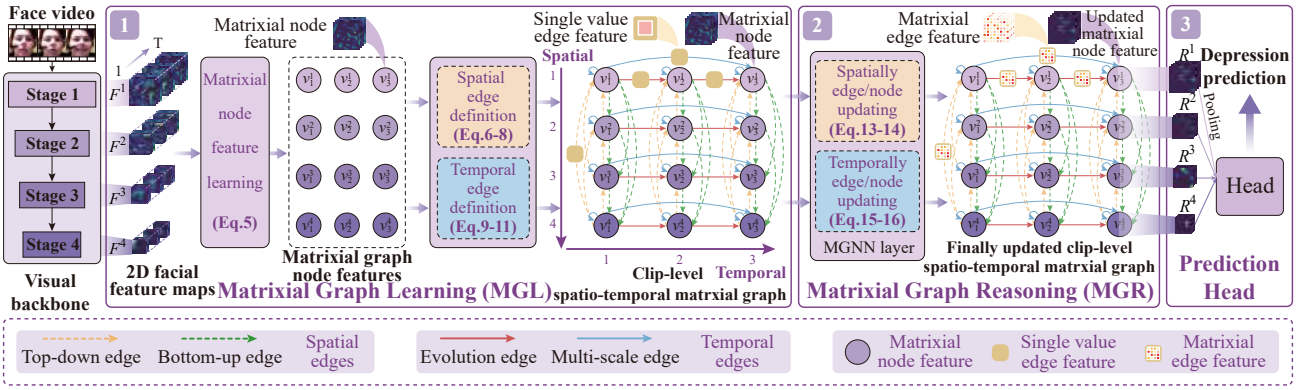

Figure 2: **Pipeline of our DepMGNN**: the backbone first extracts frame-level 2D facial feature maps at multiple spatial scales; **(1) our MGL module** learns a clip-level matrixial graph $\mathcal{G}$ to directly include the 2D facial feature maps of each frame at each spatial scale in a matrixial node v_t^n, and defines spatial and temporal edges to connect matrixial nodes; **(2)** the **MGNN** of the **MGR module** then learns a set of matrices to represent each defined edge, and conducts explicit and clip-level multi-scale spatio-temporal modelling of depression-specific facial behaviours from 2D facial feature maps (node features) of all frames; **(3)** a weighted **head** combines the updated node feature maps of all scales to make clip-level depression severity prediction.

frame-level facial features. Alternatively, spatio-temporal DL models (e.g., 3D CNNs (de Melo, Granger, and Hadid 2020; Uddin, Joolee, and Sohn 2022; Zhang et al. 2023), Temporal Pyramid Network (Xu et al. 2024), attention network (Niu et al. 2023), GNNs (Shen, Song, and Gunes 2024) and facial landmark-based temporal network (Pan et al. 2023)) also have been proposed to model segment-level depression-specific spatio-temporal facial cues from each video segment. Then, the video-level temporal information is modelled by applying LSTM (Haque et al. 2018), RNN (Al Jazaery and Guo 2018), long-term temporal differencing network (Zhang et al. 2023), eigen-evolution pooling (Niu et al. 2023) or spectral representation (Xu et al. 2024). To consider both short-term and long-term facial dynamics, alternative studies (Song et al. 2022; Song, Shen, and Valstar 2018; Jaiswal, Song, and Valstar 2019; Chen et al. 2022; Casado, Cañellas, and López 2023) encode frame-level facial features learned from all frames of the given video as a video-level spectral representation containing multi-scale facial dynamics, to make video-level depression prediction.

Vector-based GNNs: GNNs (Kipf and Welling 2016; Velickovic et al. 2017; Hamilton, Ying, and Leskovec 2017; Xu et al. 2018; Bresson and Laurent 2017) have been frequently extended to various types of video analysis tasks as well as their temporal modelling (Zeng et al. 2021b; Liu et al. 2021; Zeng et al. 2021a; Wang and Gupta 2018; Yang et al. 2020; Zhang, Tsai, and Tsai 2024). These approaches (Yang et al. 2020; Zeng et al. 2021a; Zhang, Tsai, and Tsai 2024) usually compress the spatial representations of frames/image patches into vector-style node features to accommodate existing GNNs that can only process graphs with vector-based nodes, which compromise a significant amount of spatial structural information.

3 Methodology

Overview: As illustrated in Fig. 2, our approach is made up of three modules: a **Matrixial Graph Learning (MGL)** module, a **Matrixial Graph Reasoning (MGR)** module, and a prediction head. Given a face video $X = \{I_t\}_{t\in 1,\cdots,T}$ containing T frames, our **MGL** first extracts multi-scale (i.e., N spatial scales) 2D spatial facial feature maps $f_t = \{f_t^1, \cdots, f_t^N\}$ from every frame I_t, resulting in T sets of 2D facial feature maps $F = \{f_1, \cdots, f_T\}$. Based on such frame-level feature maps, a matrixial graph $\mathcal{G}$ is constructed, each of whose node $v_t^n \in \mathcal{V}$ includes a set of 2D facial feature maps f_t^n extracted from the t_{th} frame at the n_{th} spatial scale. Meanwhile, a set of directed edges $\mathcal{E}$ are defined to connect nodes that are spatially or temporally related. Consequently, a matrixial graph $\mathcal{G}(\mathcal{V}\{F\}, \mathcal{E})$ explicitly representing multi-scale spatio-temporal dependencies among all face frames of the given video X can be obtained as:

$$\mathcal{G}(\mathcal{V}\{F\}, \mathcal{E}) = \text{MGL}(X) \tag{1}$$

As a result, variable-length face videos can be directly represented by multiple matrixial graphs containing different numbers of nodes, without re-sampling these videos, removing their frames or reducing video frames to vectors.

Subsequently, the obtained matrixial graph $\mathcal{G}(\mathcal{V}\{F\}, \mathcal{E})$ is fed to our **MGR** module, which explicitly models video-level multi-scale spatio-temporal facial cues from all frames in X (i.e., all nodes in $\mathcal{G}$). This results in a video-level depression representation $\mathcal{R} = (R^1, \cdots, R^N)$ containing N vectors describing video-level depression-specific spatio-temporal facial cues at N spatial scales as:

$$\mathcal{R} = \text{MGR}(\mathcal{G}(\mathcal{V}\{F\}, \mathcal{E})) \tag{2}$$

Finally, N learnable weights $\omega_1, \cdots, \omega_N$ are introduced to adaptively combine N vectors included in $\mathcal{R}$. The fused vector summarising multi-scale video-level spatio-tempora facial cues is then fed to a fully-connected (FC) layer $\text{FC}(\cdot)$ to

produce the video-level depression prediction P as:

$$P = \mathrm{FC}(\sum_{n=1}^{n} \omega_n R^n / (\sqrt{\sum_{n=1}^{n} {\omega_n}^2 + \varepsilon})) \tag{3}$$

where ε is a small learnable constant for numerical stability.

Training: Our entire framework is trained in a straightforward end-to-end manner, where the BMC loss (Ren et al. 2022) that addresses issues caused by unbalanced training data (i.e., depression label distribution) and the Mean Square Error (MSE) loss are joint employed with the same weight to compare the prediction y_{pred} and ground-truth y as:

$$\mathrm{L} = \underbrace{\|y - y_{\text{pred}}\|_2^2}_{\text{MSE Loss}} \underbrace{- \log \frac{\exp(-\|y - y_{\text{pred}}\|_2^2 / \tau)}{\sum_{i=1}^{B} \exp(-\|y_i - y_{\text{pred}}\|_2^2 / \tau)}}_{\text{BMC Loss}} \tag{4}$$

where $\|\cdot\|_2$ denotes the L2 norm; B represents the training batch size; and τ is a temperature coefficient that is empirically set as 2 in this paper.

3.1 Matrixial Graph Learning

The MGL module takes an arbitrary-length face video X as the input, whose CNN backbone consisting of N blocks (e.g., N residual blocks of ResNet50) learns N sets $(F^1, \cdots, F^N$, where $F^n = \{f_1^n, \cdots, f_T^n\})$ of 2D facial feature maps $f_t = \{f_t^1, \cdots, f_t^N\}$ from each frame at N different spatial scales. Based on all frame-level 2D feature maps F, a video-level facial behavioural matrixial graph $\mathcal{G}$ is constructed, where each node includes a set of matrices, while spatial edges $\mathcal{E}_S$ and temporal edges $\mathcal{E}_T$ are defined to connect spatially/temporally related nodes.

Matrix node feature learning: As shown in Fig. 2(1), given C^n 2D feature maps $f_t^n \in \mathbb{R}^{C^n \times H^n \times W^n}$ extracted from the t_{th} face frame at the n_{th} spatial scale, whose height and width are H^n and W^n, respectively, we first employ a 1×1 convolution operation $\mathrm{Conv}_{1\times1}^{(\mathrm{n})}(\cdot), n = 1, \cdots, N$ to adjust its channel number C^n (the number of frame-level 2D maps at the n_{th} spatial scale) to a fixed number (C') as:

$$f_t^{n\prime} = \mathrm{Conv}_{1\times1}^{(\mathrm{n})}(f_t^n) \tag{5}$$

where $f_t^{n\prime} \in \mathbb{R}^{C' \times H^n \times W^n}$. This ensures that the number of frame-level feature maps extracted at every spatial scale to be equal. The obtained $f_t^{n\prime}$ is then treated as a initial matrixial node feature $v_t^n \in \mathbb{R}^{C' \times H^n \times W^n}$, i.e., frame-level facial cues extracted at N spatial scales are represented by N matrixial graph nodes $\mathcal{V}_t = \{v_t^n\}_{n=1}^N$. Consequently, face videos of varying lengths can be represented by matrixial graphs containing different numbers of nodes without ignoring any frame or suppressing video frames to vectors.

Spatial edge definition: To comprehensively describe bidirectional spatial relationships among 2D feature maps extracted from different spatial scales, we define two types of directed spatial edges: top-down edges and bottom-up edges. The **top-down edges $\mathcal{E}_{\text{TD}}$** progressively connect matrixial graph nodes of larger spatial scales to nodes of smaller spatial scales belonging to the same frame as:

$$\mathcal{E}_{\text{TD}} = \{e_t^{n,m} | e_t^{n,m} = (v_t^n, v_t^m)\}_{n>m,\ n,m\in[1,N],\ t\in[1,T]} \tag{6}$$

where $e_t^{n,m}$ denotes a spatial edge connected from the node v_t^n to the node v_t^m $(n > m)$. In contrast, **bottom-up edges $\mathcal{E}_{\text{BU}}$** progressively connect nodes of smaller spatial scales to nodes of larger spatial scales of the same frame as:

$$\mathcal{E}_{\text{BU}} = \{e_t^{n,m} | e_t^{n,m} = (v_t^n, v_t^m)\}_{n<m,\ n,m\in[1,N],\ t\in[1,T]} \tag{7}$$

As a result, a set of initial spatial edges $\mathcal{E}_S = \{\mathcal{E}_{\text{TD}}, \mathcal{E}_{\text{BU}}\}$ are obtained, and an adjacency matrix A_S summarising the presence of these spatial edges is defined as:

$$\{A_S\}_t^{n,m} = \begin{cases} 1, & (v_t^n, v_t^m) \in \mathcal{E}_S \\ 0, & \text{others} \end{cases} \tag{8}$$

where $\{A_S\}_t^{n,m} \in A_S$ is an adjacency matrix component describing the presence of the spatial edge $e_t^{n,m}$ that starts from v_t^n to v_t^m. In summary, these spatial edges comprehensively facilitate both top-down and bottom-up interactions among 2D facial feature maps at N different spatial scales.

Temporal edge definition: To model temporal dependencies among all facial frames in the given video, we also define two types of directed temporal edges: evolution edges and multi-scale edges. Specifically, N sets of **evolution edges $\mathcal{E}_{\text{EV}}$** are set to model short-term temporal evolution between adjacent frames at N spatial scales, where each starts from a matrixial node v_t^n corresponding to the face frame expressed at the time t to the matrixial node v_{t+1}^n of its temporally succeeding face frame at the same spatial scale:

$$\mathcal{E}_{\text{EV}} = \{e_{t,t+1}^n | e_{t,t+1}^n = (v_t^n, v_{t+1}^n)\}_{n\in[1,N],\ t\in[1,T-1]} \tag{9}$$

Since multi-scale facial temporal information can provide complementary cues for ADA, **multi-scale temporal edges $\mathcal{E}_{\text{MT}}$** are set to connect each node v_t^n to not only all nodes of the same spatial scale $\{v_q^n\}_{q>t+1}$ extracted from facial frames expressed after the $t + 1_{\text{th}}$ frame, but also itself as:

$$\mathcal{E}_{\text{MT}} = \{e_{t,q}^n | e_{t,q}^n = (v_t^n, v_q^n)\}_{n\in[1,N],\ q=t \| q>t+1} \tag{10}$$

where $q, t \in [1, T]$. Consequently, the adjacency matrix A_T summarising these temporal edges can be defined as:

$$\{A_T\}_{t,q}^n = \begin{cases} 1, & (v_t^n, v_q^n) \in \mathcal{E}_T^n \\ 0, & \text{others} \end{cases} \tag{11}$$

where $\{A_T\}_{t,q}^n \in A_T$ describes the presence of the temporal edge $e_{t,q}^n$ that starts from the node v_t^n to the node v_q^n.

By including all spatial and temporal edges, the final adjacency matrix A describing the topology of the video-level facial behavioural matrixial graph $\mathcal{G}$ is defined as:

$$A = A_S \cup A_T \tag{12}$$

In summary, these defined edges facilitates interactions among: (i) 2D facial feature maps extracted from every frame of the given video at multiple spatial scales; and (ii) matrixial nodes corresponding to all frames of the given face video at multiple temporal scales.

3.2 Facial Matrixial Graph Reasoning

The MGR module encodes a strong depression representation from video-level depression-specific multi-scale spatio-temporal facial behaviours represented by the obtained matrixial graph $\mathcal{G}$. Specifically, we propose a novel **Matrixial Graph Neural Network (MGNN)** to facilitate a two-stage spatio-temporal modelling directly on 2D facial feature maps included in the matrixial graph $\mathcal{G}$, rather than existing GNNs (Dwivedi et al. 2023) requiring to suppress 2D feature maps as vectors beforehand. Since the binary adjacency matrix A can only define the connectivity between matrixial nodes in $\mathcal{G}$ but fail to describe their comprehensive relationships, they would prevent effective depression-specific message exchanging among matrixial nodes during the MGNN's reasoning. Specifically, given a pair of matrixical nodes v_m and v_n of the size $C \times H \times W$, theoretically there could require up to $C \times H \times W$ edge attributes (i.e., $\hat{e}_{n,m} \in \mathbb{R}^{C \times H \times W}$) to achieve $v_m = \hat{e}_{n,m} v_n$. In this sense, our MGNN innovatively learns matrix-style edge features for the spatial and temporal edges of the input matrixial graph during its reasoning (illustrated in Fig. 2 (2)).

Spatially matrixial edge and node features modelling: Given a directed spatial edge $e_t^{m,n}$ starting from the matrixial node $v_t^m \in \mathbb{R}^{C' \times H^m \times W^m}$ to the node $v_t^n \in \mathbb{R}^{C' \times H^n \times W^n}$ corresponding to the same frame t but a different spatial scale n, our MGNN deep learns a set of matrices as edge feature $\bar{e}_t^{m,n}$ which comprehensively represents spatial relationships between v_t^m to v_t^n and facilitates message passing from v_t^m to v_t^n during MGNN's reasoning as:

$$\bar{e}_t^{m,n} = \sigma(\mathrm{Conv}_{3\times3}(\mathrm{Concat}(\mathrm{resize}(v_t^m), v_t^n))) \quad (13)$$

where $\mathrm{Conv}_{3\times3}(\cdot)$ represents a 3×3 convolution operation; $\sigma(\cdot)$ is the Sigmoid function; $\mathrm{resize}(\cdot)$ denotes a downsampling or upsampling operation to adjust the shape of the source node feature v_t^m to match the shape of the destination node feature v_t^n for their fusion via the channel dimension (i.e., $\mathrm{Concat}(\cdot)$). As a result, a matrixial edge feature $\bar{e}_t^{m,n} \in \mathbb{R}^{C' \times H^n \times W^n}$ that has the same shape as the destination node v_t^n is learned. Then, our MGNN updates each node v_t^n as $\bar{v}_t^n$ by considering: (i) v_t^n itself; (ii) its spatial neighborhood $N_t(n) = \{v_t^m | \{A_\mathrm{S}\}_t^{m,n} = 1\}$ belonging to the same t_{th} frame; and (iii) the updated spatial matrixial edge features $\{\bar{e}_t^{m,n}\}_{m \in N_t(n)}$ ending at v_t^n as:

$$\bar{v}_t^n = v_t^n + \sum_{v_t^m \in N_t(n)} \bar{e}_t^{m,n} \odot \mathrm{resize}(v_t^m) \quad (14)$$

where $\odot$ represents the Hadamard product operation. This way, the updated node v_t^n aggregates multi-scale spatial facial cues of all its spatial neighbouring nodes $\{v_t^m\}_{v_t^m \in N_t(n)}$. Here, the message passing is controlled by corresponding spatial matrixial edges containing $C' \times H^n \times W^n$ deep-learned depression-specific edge attributes.

Temporally matrixial edge and node features modelling: To comprehensively model video-level depression-specific multi-scale temporal facial cues, our MGNN also learns a matrix-style feature $\hat{e}_{t,q}^n$ for each temporal edge connecting spatially updated nodes $\bar{v}_t^n$ and $\bar{v}_q^n$ which contain 2D facial feature maps extracted at the same spatial scale n but a different frame t as:

$$\hat{e}_{t,q}^n = \sigma(\mathrm{Conv}_{1\times1}^{(1)}(\bar{v}_t^n) + \mathrm{Conv}_{1\times1}^{(2)}(\bar{v}_q^n)) \quad (15)$$

The obtained $\hat{e}_{t,q}^n$ comprehensively describes temporal relationships between these two spatially updated nodes at the n_{th} spatial scale. Here, we adhere to the design paradigms of vector-based GNNs (e.g., GAT and Gated-GCN), opting not to apply normalization during node interactions but instead applying normalization after each MGNN layer. Consequently, all updated temporal matrixial edge features $\{\hat{e}_{t,q}^n\}_{n \in [0,N], q \geq t+1,\ t,q \in [1,T]}$ encompass multi-scale temporal dependencies among all frames of the given video at N spatial scales. Similar to the spatially matrixial node updating, the MGNN also passes the information from all temporal neighbouring nodes $\{\bar{v}_q^n\}_{\bar{v}_t^n \in N^n(q)}$ to the target node $\bar{v}_q^n$ via the corresponding updated temporal matrixial edges $\{\hat{e}_{t,q}^n\}_{\bar{v}_t^n \in N^n(q)}$ as:

$$\hat{v}_q^n = \bar{v}_q^n + \sum_{\bar{v}_t^n \in N^n(q)} \hat{e}_{t,q}^n \odot \bar{v}_t^n \quad (16)$$

Here, the spatio-temporal modelling from 2D facial feature maps in $\bar{v}_t^n$ to those in $\bar{v}_q^n$ is explicitly controlled by $C' \times H^n \times W^n$ depression-specific attributes in the updated temporal matrixial edge feature $\hat{e}_{t,q}^n$.

After spatially and temporally updated by (K) MGNN layers ($K = 1$ used in this paper, analysis provided in Supplementary Material), the last matrixial nodes $\hat{v}_T^n(K), n = 1, 2, \cdots, N$ of N spatial scales in the final output matrixial graph $\hat{\mathcal{G}}^K(\mathcal{V}\{F\}, \mathcal{E})$ capture video-level depression-specific multi-scale spatio-temporal facial behaviour cues. This is because in each spatial scale, all nodes are connected to the last node, and pass their information to it via their temporal edges. As a result, our MGR encodes matrices in these N nodes as N vectors $\mathcal{R} = (R^1, \cdots, R^N)$ via the global average pooling, for the final depression prediction.

4 Experiment

4.1 Experimental Setup

Datasets. We evaluate our approach on two visual depression assessment benchmark datasets: AVEC 2013 (Valstar et al. 2013) and AVEC 2014 (Valstar et al. 2014). Participants in both datasets are labeled with depression levels ranging from 0 to 63 based on the Beck Depression Inventory (BDI-II). The AVEC 2013 recorded a total of 150 video clips ranging from 20 to 50 mins from 82 participants. The AVEC 2014 involves two types of recording conditions: Northwind and Freeform, which contains 300 clips with each subset containing 150 clips ranging from 6 seconds to 4 mins 8 seconds. Both datasets were evenly divided into three subsets: training, development and testing.

Metrics. We follow previous studies (Valstar et al. 2013, 2014; Ringeval et al. 2019) to compare our approach with existing methods to measure by measuring errors and correlations between predictions and ground-truths using: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) Pearson Correlation Coefficient (PCC) and Concordance Correlation Coefficient (CCC).

Implementation details. The face region of each video frame is first cropped and resized to 224×224. Then, we choose ResNet50 (He et al. 2016) and SENet (Hu, Shen, and Sun 2018) pre-trained on VGGFace2 (Cao et al. 2018) as our backbones. For each dataset, we develop models on training and development sets, while report the results achieved on the test set. More details of the employed datasets, implementation (hyper-parameters, libraries and hardware) and metrics are provided in Supplementary Material.

4.2 Comparison to Existing Approaches

Table 1 demonstrates that our approach clearly outperformed all existing competitors, and achieved new state-of-the-art (SOTA) performances on both datasets with 2.5% and 5.2% relative RMSE improvements on AVEC2013 and AVEC2014 datasets over previous SOTA methods. Compared to recently proposed graph-based ADA method (Xu et al. 2024), our approach also shows more than 6.5% and 13.5% relative RMSE improvements on two datasets, emphasizing the advantage of the proposed matrixial video graph representations and MGNN over standard video graph representations and GNNs in modelling depression-specific spatio-temporal facial behaviours. In AVEC2014 dataset, applying our approach to individually model task-dependent behaviours (expressed via FreeForm and Northwind tasks) and fusing their predictions via one layer MLP (F-N fusion) further improved the performance, suggesting that our MGNN can learn informative and complementary depression-specific facial cues from behaviours triggered by different stimulus. To further evaluate the generalization capability of our approach, we additionally apply our approach to video-based personality recognition that aims to predict five personality traits from each video. The results provided in Supplementary Material show that our matrixial graph and MGNN also achieved new SOTA results for this task.

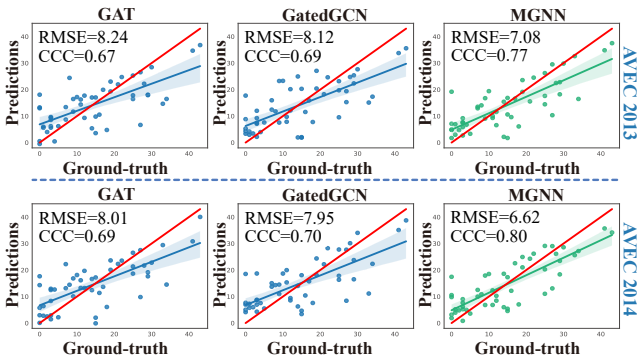

Figure 3: Visual comparison among predictions achieved by our matrixial face video graphs containing matrix-style node features (processed by our MGNN) and face video graphs containing vector-based node features (processed by GAT and GatedGCN) on AVEC 2013 and AVEC 2014 datasets.

4.3 Ablation Studies

To deeply investigate the effectiveness of our approach, we conduct a series of ablation studies on AVEC 2014 dataset as below. Due to the limited space, we additionally provide: (1) statistical difference; (2) cross-datasets evaluation; (3) model complexity and running time; (4) more visualization results; (5) the number (K) of MGNN layers; as well as the results achieved for: (6) video segments to validate the importance of modelling video-level cues; (7) different loss settings; and (8) other analysis in **Supplementary Material**.

Methods	AVEC 2013		AVEC 2014	
	RMSE↓	MAE↓	RMSE↓	MAE↓
(Jan et al. 2017)	-	-	8.01	6.68
(Meng et al. 2013)	11.19	9.14	-	-
(He, Jiang, and Sahli 2018)	9.20	7.55	9.01	7.21
(Niu et al. 2020)	8.97	7.32	8.60	6.43
(Uddin, Joolee, and Lee 2020)	8.93	7.04	8.78	6.86
(He, Chan, and Wang 2021)	8.39	6.59	8.30	6.51
(Zhou et al. 2020)	8.37	6.63	8.30	6.59
(Zhou et al. 2018)	8.28	6.20	8.39	6.21
(de Melo, Granger, and Hadid 2019)	8.26	6.40	8.31	6.59
(Shang et al. 2021)	8.20	6.38	7.84	6.08
(Song et al. 2022)	8.10	6.16	7.15	5.95
(Niu et al. 2020)	7.98	6.15	7.75	6.00
(de Melo, Granger, and Hadid 2020)	7.90	5.98	7.61	5.82
(Liu et al. 2023)	7.59	6.08	7.98	6.04
(Xu et al. 2024)	7.57	5.95	7.65	6.24
(de Melo, Granger, and Lopez 2021)	7.55	6.24	7.65	6.06
(Niu et al. 2022a)	7.49	6.12	7.56	6.01
(Uddin, Joolee, and Sohn 2022)	7.32	5.90	6.98	5.75
(Pan et al. 2023)	[7.26]	5.97	7.30	5.99
Ours (SENet-50)	[7.26]	**5.38**	6.76	[5.45]
Ours (ResNet-50)	**7.08**	[5.41]	[6.62]	[5.45]
Ours (F-N fusion)	-	-	**6.28**	**4.99**

Table 1: Comparison with other video-based ADA methods, where bold and bracketed values denote the best and second best results. The F-N fusion denotes conducting decision-level fusion from Freefrom and Northwind videos to make each individual-level prediction (also conducted by other competitors, e.g., (Song et al. 2022; Xu et al. 2024)).

Contributions of the MGL and MGR/MGNN: Table 2 indicates that matrixial graph representations learned by our **MGL** can effectively model depression-specific facial cues from videos (even reducing the learned matrixial nodes to vector-based nodes and processed by widely-used GAT (Veličković et al. 2017) and GatedGCN (Bresson and Laurent 2017)), resulting in clear performance gains. Based on the same matrixial graph learned by **MGL**, our **MGNN** is clearly superior to GAT and GatedGCN that can only process vector-based node features, despite that MGNN and GatedGCN share the same node/edge feature updating mechanisms (also compared in Fig. 3). Besides, Table 3 and Fig. 4 additionally show the advantages of our MGNN over other models that can handle variable-length data, where our MGNN: (1) does not require to suppress facial feature maps to vector-based representations as LSTM/GRU/GAT/GatedGCN; and (2) can batch process graphs containing different numbers of nodes (i.e., batch processing videos of varying lengths), while the batch processing of /LSTM/GRU/Conv3D/Transformers (e.g. ViT (Dosovitskiy et al. 2020) and Timesformer (Bertasius, Wang, and Torresani 2021)) require all examples within a batch have the same size. *The above results indicate that compared to suppressing face frames as vector-based features, our strategy which directly models 2D facial feature maps via matrix-style edge features can retain and extract more depression-*

Backbone	MGL		GNN				Head	RMSE↓	MAE↓	PCC↑	CCC↑
	Spatial	Temporal	GAT	GatedGCN	MGNN(s)	MGNN(m)					
✓							✓	8.44	6.49	0.68	0.66
✓	✓	✓	✓				✓	8.01	6.21	0.72	0.69
✓	✓	✓		✓			✓	7.95	6.39	0.72	0.70
✓	✓	✓			✓		✓	7.54	6.07	0.76	0.75
✓	✓					✓	✓	7.49	5.87	0.76	0.74
✓		✓				✓	✓	7.57	5.91	0.76	0.73
✓	✓	✓				✓		6.79	5.53	0.81	0.79
✓	✓	✓				✓	✓	**6.62**	**5.45**	**0.82**	**0.80**

Table 2: Results achieved for different model settings. When standard GAT and GatedGCN is applied, all matrixial node features are projected to vector-based node features. For the system without 'head', they simply averages predictions achieved from all spatial scales. (s) and (m) denote the adjacency matrix-based binary-value edges and matrixial edge features, respectively.

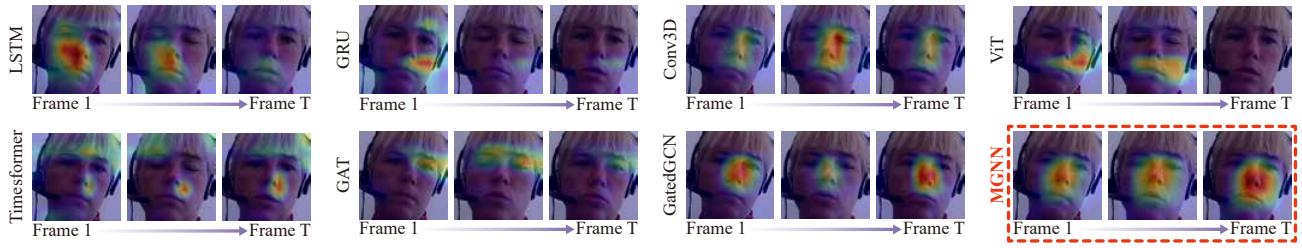

Figure 4: Latent feature visualization of systems that replace MGNN with other layers when processing our matrixial graphs, where 3DConv/ViT/TimesFormer and our MGNN directly process matrixial node features (2D facial feature maps) while others suppressed matrixial node features as vectors beforehand. Compared to other widely-used layers, our MGNN consistently utilized cues from more face regions for depression assessment, validating its effectiveness in spatio-temporal modelling.

specific spatio-temporal facial cues.

Method	RMSE↓	MAE↓	PCC↑	CCC↑
LSTM	8.55	6.82	0.67	0.65
GRU	8.40	6.63	0.69	0.65
Conv3D	8.78	7.10	0.65	0.57
ViT	8.54	6.61	0.68	0.66
Timesformer	8.10	6.51	0.71	0.66
Ours (MGNN)	**6.62**	**5.45**	**0.82**	**0.80**

Table 3: Results of replacing MGNN with other layers.

Contributions of spatial and temporal edges: Table 2 also shows that multi-scale message exchanging via both spatial and temporal edges with matrixial edge features are important for video-based depression assessment, with spatial interactions achieved a slightly better performance. Here, our novel matrixial edge features bought significantly benefits over regular binary-value edges. Importantly, these edges allow our approach to achieve large improvements over the backbone that does not facilitate multi-scale spatial or temporal interactions during feature extraction. These results validates not only ***the importance of conducting spatial and temporal modelling during feature extraction but also the effectiveness of our novel matrix-style edge features in matrixial node features interactions.***

Contributions of different spatial scales: Table 4 compares the results achieved by applying MGNN to process matrixial graphs that encode depression-specific facial cues at different spatial scales. It is clear that considering facial behaviours at more spatial scales are beneficial, where the performance are consistently enhanced when considering more spatial scales. Then, we found that facial behaviours at every spatial scale can partially reflect depression status (evidenced by their high CCC performances), which are complementary to each other as their combination always outperformed variants only considered one spatial scale.

MGNN				RMSE↓	MAE↓	PCC↑	CCC↑
S1	S2	S3	S4				
✓				9.25	7.17	0.60	0.57
	✓			8.89	7.16	0.65	0.64
		✓		**8.69**	**6.59**	**0.67**	**0.65**
			✓	8.78	7.10	0.65	0.61
✓	✓			9.03	7.22	0.64	0.62
✓		✓		8.41	6.80	0.69	0.67
✓			✓	**7.99**	**6.26**	**0.75**	**0.72**
	✓	✓		8.43	6.65	0.69	0.65
	✓		✓	8.29	6.81	0.71	0.68
		✓	✓	8.12	6.45	0.72	0.71
✓	✓	✓		8.07	6.24	0.72	0.70
✓	✓		✓	7.57	**5.74**	0.76	0.74
✓		✓	✓	7.33	5.92	0.74	0.73
	✓	✓	✓	**7.19**	5.83	**0.78**	**0.76**
✓	✓	✓	✓	**6.62**	**5.45**	**0.82**	**0.80**

Table 4: Results achieved for different spatial scales.

Discussion: According to ablation studies reported above and the Supplementary Material, we found that: (1) both **MGL** and **MGR** modules are effective; (2) our novel matrixial edge features and multi-scale spatio-temporal mod-

elling are crucial; (3) our approach generalize well for cross-dataset evaluation; (4) our novel MGNN is superior to transformer and LSTM in handling variable-length videos for ADA; and (5) modelling facial cues at the long-term video-level is necessary and beneficial.

5 Conclusion

This paper proposes a novel video-based ADA approach, where our proposed novel matrixial graph and MGNN have clear advantages over graphs consisting of vector-based node features and standard GNNs for this task, making our approach as the new SOTA video-based ADA solution. However, this effectiveness is at the cost of the increased model complexity, and thus more efficient training strategies is our future research direction.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant No. 62271263, U23A20283, 62306068), and the Natural Science Foundation of Hebei Province, China (Grant No. F2024501002).

References

Al Hanai, T.; Ghassemi, M. M.; and Glass, J. R. 2018. Detecting Depression with Audio/Text Sequence Modeling of Interviews. In *Interspeech*, 1716–1720.

Al Jazaery, M.; and Guo, G. 2018. Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing*.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, 4.

Bresson, X.; and Laurent, T. 2017. Residual gated graph convnets. *arXiv preprint arXiv:1711.07553*.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.

Casado, C. Á.; Cañellas, M. L.; and López, M. B. 2023. Depression recognition using remote photoplethysmography from facial videos. *IEEE Transactions on Affective Computing*.

Chen, M.; Xiao, X.; Zhang, B.; Liu, X.; and Lu, R. 2022. Neural Architecture Searching for Facial Attributes-based Depression Recognition. *arXiv preprint arXiv:2201.09799*.

Clark, L. A.; and Watson, D. 1991. Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *Journal of abnormal psychology*, 100(3): 316.

de Melo, W. C.; Granger, E.; and Hadid, A. 2019. 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)*, 1–8. IEEE.

de Melo, W. C.; Granger, E.; and Hadid, A. 2020. A deep multiscale spatiotemporal network for assessing depression from facial dynamics. *IEEE Transactions on Affective Computing*.

de Melo, W. C.; Granger, E.; and Lopez, M. B. 2021. Mdn: A deep maximization-differentiation network for spatio-temporal depression detection. *IEEE Transactions on Affective Computing*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dwivedi, V. P.; Joshi, C. K.; Luu, A. T.; Laurent, T.; Bengio, Y.; and Bresson, X. 2023. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24(43): 1–48.

Ellgring, H. 2007. *Non-verbal communication in depression*. Cambridge University Press.

Fang, M.; Peng, S.; Liang, Y.; Hung, C.-C.; and Liu, S. 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82: 104561.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Haque, A.; Guo, M.; Miner, A. S.; and Fei-Fei, L. 2018. Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions. *arXiv preprint arXiv:1811.08592*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, L.; Chan, J. C.-W.; and Wang, Z. 2021. Automatic depression recognition using CNN with attention mechanism from videos. *Neurocomputing*, 422: 165–175.

He, L.; Jiang, D.; and Sahli, H. 2018. Automatic Depression Analysis using Dynamic Facial Appearance Descriptor and Dirichlet Process Fisher Encoding. *IEEE Transactions on Multimedia*.

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

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jaiswal, S.; Song, S.; and Valstar, M. 2019. Automatic prediction of Depression and Anxiety from behaviour and personality attributes. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–7.

James, S. L.; Abate, D.; Abate, K. H.; Abay, S. M.; Abbafati, C.; Abbasi, N.; Abbastabar, H.; Abd-Allah, F.; Abdela, J.; Abdelalim, A.; et al. 2018. 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.

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

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Liu, J.; Rojas, J.; Li, Y.; Liang, Z.; Guan, Y.; Xi, N.; and Zhu, H. 2021. A graph attention spatio-temporal convolutional network for 3D human pose estimation in video. In *2021 IEEE international conference on robotics and automation (ICRA)*, 3374–3380. IEEE.

Liu, Z.; Yuan, X.; Li, Y.; Shangguan, Z.; Zhou, L.; and Hu, B. 2023. PRA-Net: Part-and-Relation Attention Network for depression recognition from facial expression. *Computers in Biology and Medicine*, 157: 106589.

Meng, H.; Huang, D.; Wang, H.; Yang, H.; Ai-Shuraifi, M.; and 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*, 21–30.

Niu, M.; He, L.; Li, Y.; and Liu, B. 2022a. Depressioner: Facial dynamic representation for automatic depression level prediction. *Expert Systems with Applications*, 204: 117512.

Niu, M.; Tao, J.; Liu, B.; Huang, J.; and Lian, Z. 2020. Multimodal Spatiotemporal Representation for Automatic Depression Level Detection. *IEEE Transactions on Affective Computing*.

Niu, M.; Tao, J.; Liu, B.; Huang, J.; and Lian, Z. 2023. Multimodal Spatiotemporal Representation for Automatic Depression Level Detection. *IEEE Transactions on Affective Computing*, 14(01): 294–307.

Niu, M.; Zhao, Z.; Tao, J.; Li, Y.; and Schuller, B. W. 2022b. Dual attention and element recalibration networks for automatic depression level prediction. *IEEE Transactions on Affective Computing*, 14(3): 1954–1965.

Pampouchidou, A.; Simantiraki, O.; Fazlollahi, A.; Pediaditis, M.; Manousos, D.; Roniotis, A.; Giannakakis, G.; Meriaudeau, F.; Simos, P.; Marias, K.; et al. 2016. Depression assessment by fusing high and low level features from audio, video, and text. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 27–34.

Pan, Y.; Shang, Y.; Liu, T.; Shao, Z.; Guo, G.; Ding, H.; and Hu, Q. 2024. Spatial–temporal attention network for depression recognition from facial videos. *Expert Systems with Applications*, 237: 121410.

Pan, Y.; Shang, Y.; Shao, Z.; Liu, T.; Guo, G.; and Ding, H. 2023. Integrating deep facial priors into landmarks for privacy preserving multimodal depression recognition. *IEEE Transactions on Affective Computing*.

Ray, A.; Kumar, S.; Reddy, R.; Mukherjee, P.; and Garg, R. 2019. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, 81–88.

Ren, J.; Zhang, M.; Yu, C.; and Liu, Z. 2022. Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7926–7935.

Ringeval, F.; Schuller, B.; Valstar, M.; Cummins, N.; Cowie, R.; Tavabi, L.; Schmitt, M.; Alisamir, S.; Amiriparian, S.; Messner, E.-M.; et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, 3–12.

Ringeval, F.; Schuller, B.; Valstar, M.; Gratch, J.; Cowie, R.; Scherer, S.; Mozgai, S.; Cummins, N.; Schmitt, M.; and Pantic, M. 2017. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 3–9.

Shang, Y.; Pan, Y.; Jiang, X.; Shao, Z.; Guo, G.; Liu, T.; and Ding, H. 2021. LQGDNet: A local quaternion and global deep network for facial depression recognition. *IEEE Transactions on Affective Computing*, 14(3): 2557–2563.

Shen, H.; Song, S.; and Gunes, H. 2024. Multi-modal Human Behaviour Graph Representation Learning for Automatic Depression Assessment.

Song, S.; Jaiswal, S.; Shen, L.; and Valstar, M. 2022. Spectral Representation of Behaviour Primitives for Depression Analysis. *IEEE Transactions on Affective Computing*, 13(2): 829–844.

Song, S.; Luo, Y.; Tumer, T.; Fu, C.; Valstar, M.; and Gunes, H. 2024. Loss relaxation strategy for noisy facial video-based automatic depression recognition. *ACM Transactions on Computing for Healthcare*, 5(2): 1–24.

Song, S.; Shen, L.; and Valstar, M. 2018. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 158–165. IEEE.

Uddin, M. A.; Joolee, J. B.; and Lee, Y.-K. 2020. Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm. *IEEE Transactions on Affective Computing*.

Uddin, M. A.; Joolee, J. B.; and Sohn, K.-A. 2022. Deep multi-modal network based automated depression severity estimation. *IEEE transactions on affective computing*, 14(3): 2153–2167.

Valstar, M.; Schuller, B.; Smith, K.; Almaev, T.; Eyben, F.; Krajewski, J.; Cowie, R.; and Pantic, M. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, 3–10.

Valstar, M.; Schuller, B.; Smith, K.; Eyben, F.; Jiang, B.; Bilakhia, S.; Schnieder, S.; Cowie, R.; and 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, 3–10.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wang, X.; and Gupta, A. 2018. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, 399–417.

Xu, J.; Gunes, H.; Kusumam, K.; Valstar, M.; and Song, S. 2024. Two-stage temporal modelling framework for video-based depression recognition using graph representation. *IEEE Transactions on Affective Computing*.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Yang, J.; Zheng, W.-S.; Yang, Q.; Chen, Y.-C.; and Tian, Q. 2020. Spatial-temporal graph convolutional network for video-based person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3289–3299.

Zeng, R.; Huang, W.; Tan, M.; Rong, Y.; Zhao, P.; Huang, J.; and Gan, C. 2021a. Graph convolutional module for temporal action localization in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6209–6223.

Zeng, X.; Jiang, Y.; Ding, W.; Li, H.; Hao, Y.; and Qiu, Z. 2021b. A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1): 200–212.

Zhang, J.; Tsai, P.-H.; and Tsai, M.-H. 2024. Semantic2Graph: graph-based multi-modal feature fusion for action segmentation in videos. *Applied Intelligence*, 1–16.

Zhang, S.; Zhang, X.; Zhao, X.; Fang, J.; Niu, M.; Zhao, Z.; Yu, J.; and Tian, Q. 2023. MTDAN: A lightweight multi-scale temporal difference attention networks for automated video depression detection. *IEEE Transactions on Affective Computing*.

Zhou, J.; Zhang, X.; Liu, Y.; and Lan, X. 2020. Facial expression recognition using spatial-temporal semantic graph network. In *2020 IEEE International Conference on Image Processing (ICIP)*, 1961–1965. IEEE.

Zhou, X.; Jin, K.; Shang, Y.; and Guo, G. 2018. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*.