Explainable Depression Assessment from Face Videos by Weakly Supervised Learning

Rongfan Liao[1,2], Xiangyu Kong[2], Shiqing Tang[2,3], Lang He[4], Changzeng Fu[5], Weichang Xie[6], Xiaofeng Liu[7], Lu Liu[2], Siyang Song[2*]

[1]University of Leicester, UK
[2]HBUG Lab, University of Exeter, UK
[3]Shanghai University of Science and Technology, China
[4]Xi'an University of Posts and Telecommunications, China
[5]Northeastern University, China
[6]Shenzhen University, China
[7]Hohai University, China

Abstract

Existing video-based automatic depression assessment (ADA) approaches frequently achieve video-level depression assessment by aggregating features or predictions of individual frames or equal-length segments within the given video. While their performances have been largely enhanced by recent advanced deep learning models, they typically fail to explicitly consider the varied importance of depression-related behavioural cues across different video segments, i.e., segments within one video may contain behaviours reflecting varying levels of depression. Underestimating segment-level variations can obscure the detection of facial behaviour cues associated with depression, thereby undermining the accuracy and interpretability of video-based depression detection systems. In this paper, we propose a novel video-based ADA approach that specifically identifies and differentiates video segments that exhibit depression-related facial behaviours across varying temporal durations, providing clear insights into how each segment contributes to the video-level depression prediction. To achieve this, a novel weakly supervised strategy is proposed to compare segment-level behaviours with video-level depression label, enabling the model to assign depression-relevant scores to multiple temporal scale video segments and attend selectively to those most indicative of depressive states. Extensive experiments on the AVEC 2013 and AVEC 2014 face video depression datasets demonstrate the effectiveness of our approach.

Code — https://github.com/liaorongfan/ExpADA

Introduction

Depression is the most common mental health disorder characterised by persistent negative emotions and a loss of interest in activities (Rottenberg 2017). It manifests through a range of symptoms that affect human emotion, cognitive processes, and physical well-beings, sometimes even escalating to suicidal thoughts and behaviours (Stehman et al. 2019; Zhang et al. 2020). To prevent the progression of depression, early detection plays an important role in facilitating timely and proper treatment. While previous studies have consistently shown that facial behaviours provide reliable indicators of depression (Ellgring 2007), which can

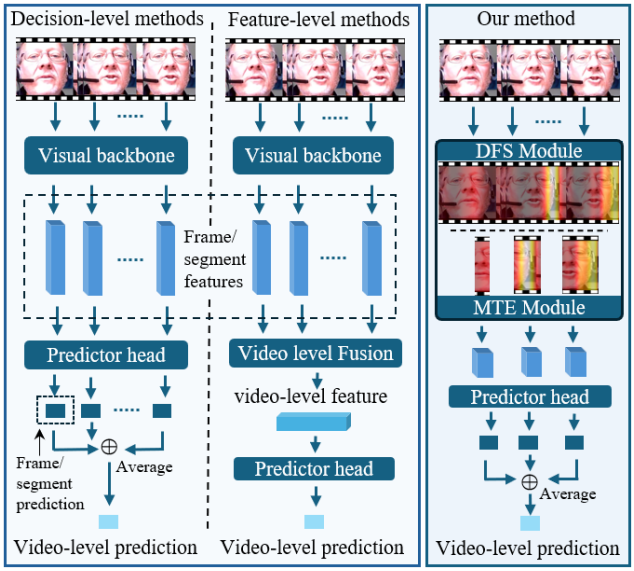

Figure 1: Current video-based ADA typically employs **decision-level methods** that average frame/segment-level predictions into video-level prediction or **feature-level methods** that fuse features extracted from frames/segments to model video-level temporal representation for depression prediction. However, decision-level methods take static frames or equal-length segments as input, which fall short in capturing depression-related facial behaviours across various durations. Moreover, both decision- and feature-level methods normally treat every frame/segment equally in their relevance to the video-level depression severity, which overlook the negative influence of less relevant or noisy frames/segments within the video. **Our method** selects depression-relate segments at various temporal scales for video level depression detection, which not only improves the accuracy but also provide interpretability for the video-base ADA process.

*Corresponding author (s.song@exeter.ac.uk).
This is the author's version of the paper accepted for publication at AAAI 2026.

be easily captured via non-invasive cameras, numerous face video-based automatic depression assessment (ADA) methods have been proposed (Song et al. 2022; Jaiswal, Song, and Valstar 2019; De Melo, Granger, and Lopez 2020; Uddin, Joolee, and Lee 2020; Song et al. 2024), to offer easily accessible ways for daily depression monitoring.

Recent advances in deep learning (DL) have demonstrated significant potentials in modeling depression-related spatio-temporal facial cues for video-based ADA (He et al. 2022; Niu et al. 2022; de Melo, Granger, and Lopez 2021). Due to the high memory costs in directly processing long videos, most of these DL-based approaches divide videos into a set of individual frames or equal-length short segments (Liao, Song, and Gunes 2024). These approaches typically model depression-relevant representations (Xu et al. 2024; Niu et al. 2020) or make predictions (Mao et al. 2022; De Melo, Granger, and Lopez 2020; Niu et al. 2022) from static-frames/short-segments, and then aggregate the representations/predictions to achieve the final video-level depression prediction. They equally associate every frame/short segment with the corresponding overall video-level depression label to train their models, despite that the depression diagnostic relevance of facial behaviours can vary substantially across different frames/segments of the same video (He et al. 2022). As a result, their training may be degraded by pairing facial frames/segments with incorrect depression labels, impairing their ability to detect the most indicative facial behaviours associated with depression.

Due to the lack of well-annotated labels to assess the relevance of each face video frame/segment to its video-level depression severity, existing video-based ADA methods typically fail to accurately differentiate the importance of every frame/segment contributing to final video-level depression status prediction nor prioritise the most depression-relevant facial behaviours for the final assessment (**Problem 1**). Although depression is more reliable to be reflected by long-term facial behaviours (Wang et al. 2024), only a few of existing methods (Gahalawat et al. 2023; Pan et al. 2023a; He et al. 2024c) explain frame-level or short segment-level depression-related facial behaviours, leaving multi-scale video-level interpretability underexplored. Moreover, the depression-related facial behaviours could vary in their duration, making static-frame-based and fixed-length-segment based methods fail to accurately explain all crucial facial-behavioural depression biomarkers (**Problem 2**). Addressing these problems can not only facilitate to train reliable video-level ADA models by locating depression-related facial behaviours while minimising the impact of irrelevant facial behaviours, but also enhance interpretability of the ADA process.

This paper presents a novel video-based explainable ADA approach that identifies depression-related facial behaviours by analysing the relevance of video segments at multiple temporal scales to video-level depression status. Specifically, the approach begins by extracting spatio-temporal features from equal-length video segments, which are then sequentially combined to form facial behaviour representations at multiple temporal scales. These representations are subsequently scored to reflect their relevance to the final video-level depression prediction, enabling the model to prioritise the most informative facial behaviours of various temporal scales in deciding the final video-level depression severity. Since only video-level depression labels are available for the model training, we propose a novel weakly supervised training strategy that selects facial behaviours within multiple temporal scale segments which are associated with the labelled video-level depression severity category (none, mild, modest or severe). This way, our approach provides a more accurate and consistent analysis for video-level depression assessment. Figure 1 compares our novel approach with existing video-based ADA solutions. The main novelties and contributions of this paper are summarised as follows:

- We propose a novel weakly supervised training strategy that specifically identifies depression-informative facial behaviours across multiple temporal scales in face videos, which enables the selection of the most relevant facial segments at multiple temporal scales for more precise video-based ADA, while mitigating the influence of unrelated or noisy segments.
- We propose a novel video-level explainable ADA approach that identifies and localizes depression-relevant facial behaviours across multiple temporal scales. To the best of our knowledge, this is the first attempt to provide video-level temporally fine-grained interpretability for video-based ADA at multiple temporal scales.
- Extensive experiments demonstrate that our approach achieved state-of-the-art performance in video-based ADA on AVEC 2013 and 2014 benchmark datasets. Especially on AVEC 2014 dataset, our approach outperforms most existing models in the field.

Related Work

Video-based automatic depression assessment: Due to the limited memory and computation capability of existing devices, most existing video-based ADA approaches extract depression-relevant information either at the frame level or from uniformly divided video segments. Specifically, frame-level ADA methods (Uddin, Joolee, and Lee 2020; He, Chan, and Wang 2021; Yang et al. 2024; Pan et al. 2024) frequently learn static facial features from every frame, capturing frame-level static facial cues associated with depression. Meanwhile, segment-level methods (Zhou et al. 2020; Zhang et al. 2023; He et al. 2024a; Xu et al. 2024) divide videos into a set of fixed-length segments and extract short-term spatio-temporal facial features from each of them. Subsequently, some studies (Niu et al. 2020; Yan et al. 2024; He et al. 2025) propose various strategies to effectively integrate (e.g., temporal modelling) features/predictions across all frame/segment-level features for more accurate and robust video-level ADA. To capture more holistic video-level information, some alternative ADA approaches (Song et al. 2022; de Melo, Granger, and Lopez 2021; Wu et al. 2025; Song, Shen, and Valstar 2018) directly construct video-level representation to infer depression, despite that they fail to include all facial behaviours provided by the video as they remove frames/spectral components during the video-level

representation learning. More importantly, none of such approaches specifically considers the varying differences of multi-scale frame/segment-level facial behaviours in reflecting video-level depression severity, thereby suffering from poor interpretability of their predictions.

Explainable video-base depression assessment: There a few prior approaches attempted to explain video-based ADA results in terms of frame-level or short-term facial behaviours. Some of them apply a set of pre-defined facial primitives to provide explanations. For example, based on facial action units (AUs), head pose, and gaze trajectories, Mahayossanunt et al. (Mahayossanunt et al. 2023) employs Integrated Gradients (Sundararajan, Taly, and Yan 2017) to quantify the importance of each descriptor in predicting PHQ scores. Gahalawat et al. (Gahalawat et al. 2023) models head movements using a set of discrete motion primitives (e.g., nod, shake, tilt), revealing that depressed participants often lack typical head gestures like upward nods. Niu et al. (Niu, Li, and Fu 2024) leverages self-attention to highlight most informative spatio-temporal facial behaviours in the form of their facial landmark sequence. Alternatively, other solutions aim to explain the prediction at the pixel level. The STA-DRN (Pan et al. 2023a) employs spatial-temporal attention maps to highlight depression-discriminative facial regions such as the eye and cheek areas. The TCEDN (Yan et al. 2024) enhances interpretability through attention-based feature aggregation and motion encoding. Grad-CAM visualisation on input images consistently attends to clinically relevant regions, such as the eyes, eyebrows, and surrounding areas. Similarly, He et al. (He et al. 2024c) stacks multi-scale Transformer blocks on a 3-D CNN backbone to produce frame-level Grad-CAM maps across every short facial behaviour segments, suggesting that eye blinks, mouth droops and cheek twitches are crucial for ADA.

Methodology

Overview: As illustrated in Figure 2, given a face video V that has been divided into N non-overlapped equal-length segments $V = \{v_n\}_{n=1}^{N}$, our approach starts with the **Multi-scale Temporal Behaviour Encoder (MTE)** which first encodes each segment v_n into a segment-level facial behaviour feature unit (FU) $\mathbf{f}_n^v \in \mathbb{R}^D$ (D denotes the feature dimension). Then, it processes all FUs ($\{\mathbf{f}_n^v\}_{n=1}^{N}$) extracted from the given video V into S sets of representations $\mathcal{F} = \{F^{(s)}\}_{s\in\{1,\dots,S\}}$ describing facial behaviours from S different temporal scales as:

$$\mathcal{F} = \text{MTE}(V), \tag{1}$$

where $F^{(s)} = \{\mathbf{f}_n^{(s)}\}_{n=1}^{N-s+1} \in \mathcal{F}$ corresponds to the representation set at the s temporal scale, then the total number of representations in $\mathcal{F}$ can be denoted as $M = \sum_{s\in\mathcal{S}}(N - s + 1)$.

Subsequently, a **Discriminative Feature Selection (DFS)** module selects depression-informative facial-behaviour representations at multiple temporal scales from $\mathcal{F}$ for final depression score prediction as:

$$\mathcal{F}_{\text{sel}} = \text{DFS}\big(\mathcal{F}\big), \qquad \mathcal{F}_{\text{sel}} \subset \mathcal{F}. \tag{2}$$

Specifically, for each facial-behaviour representation $\mathbf{f}_n^{(s)} \in \mathcal{F}$ extracted across temporal scale s, The DFS module predicts a raw depression category and its confidence:

$$(\hat{y}_n^{(s)}, \hat{c}_n^{(s)}) = \text{DFS}(\mathbf{f}_n^{(s)}), \quad \forall \mathbf{f}_n^{(s)} \in \mathcal{F} \tag{3}$$

where $\hat{y}_n^{(s)} \in \{\text{none}, \text{mild}, \text{moderate}, \text{severe}\}$ and $\hat{c}_n^{(s)} \in [0, 1]$. The video-level depression category is decided by the most frequently predicted category among all the representations in $\mathcal{F}$ as:

$$I = \text{mode}\big\{\hat{y}_n^{(s)} \mid \forall s, n\big\}, \tag{4}$$

Then, the DFS module selects the top $p\%$ most confident representations that support the video-level depression category I, while ignoring the other less relevant representations as:

$$\mathcal{F}_{\text{sel}} = \text{Top}_{p\%}\big(\{\mathbf{f}_n^{(s)} \mid \hat{y}_n^{(s)} = I\}, \hat{c}_n^{(s)}\big). \tag{5}$$

Finally, a predictor head processes the selected representations $\mathcal{F}_{\text{sel}}$ to predict individual depression scores, which are then averaged into the final video-level depression severity score, as:

$$D_s = \text{mean}(\text{FC}(\mathcal{F}_{\text{sel}})) \tag{6}$$

where $FC(\cdot)$ represents fully connected layers.

Depression Indicative Segments Identification

Multi-scale Temporal Behaviour Encoder The MTE module follows previous video-based ADA approaches (Xu et al. 2024; He et al. 2024b) to first divide the given video V into a sequence of equal-length (T frames) segments $\{v_n\}_{n=1}^{N}$, followed by a Video Action Transformer (VAT) (Girdhar et al. 2019) employed to extract a segment-level facial behaviour feature $\mathbf{f}_n^v$ (referred as a facial behaviour feature unit (FU)) from each segment $\{v_n\}$. Then, we propose a multi-scale feature fusion strategy on the obtained FUs to capture facial behaviours at multiple temporal scales. As illustrated in Figure 2, the MTE applies a sliding window mechanism over the sequence of extracted FUs $\{\mathbf{f}_n^v\}_{n=1}^{N}$ to construct facial behaviour representations at S different temporal scales by concatenating s ($s = \{1, 2, \cdots, S | S \ll N\}$) consecutive FUs in the temporal dimension with a stride of one FU. These concatenated representations describing facial behaviours at S different temporal scales are individually fed to S distinct Multi-Layer Perceptrons (MLPs), projecting them into a latent space with the same dimension D as:

$$F^{(s)} = \{\mathbf{f}_n^{(s)}\}_{n=1}^{N-s+1} \quad \mathbf{f}_n^{(s)} = \text{MLP}_s(\text{cat}(\{\mathbf{f}_i^v\}_{i=n}^{n+s-1})) \in \mathbb{R}^D \tag{7}$$

where $\text{cat}(\cdot)$ denotes concatenation operation. This way, representation sets $\mathcal{F} = \{F^{(s)}\}_{s\in\{1,\dots,S\}}$ describing facial behaviours at S different temporal scales can be obtained.

Discriminative Feature Selection As illustrated in Figure 2, every temporal representation $\mathbf{f}_n^{(s)} \in \mathbb{R}^D$ in the obtained $\mathcal{F}$ is first fed to a MLP-based classification branch to individually predict its indicative depression category as:

$$\mathbf{P}_{n,c} = \underset{n}{\text{Softmax}}(\hat{\mathbf{H}}_1) \quad \hat{\mathbf{H}}_1 = \text{MLP}_{\text{rel}}(\{\mathbf{f}_n^{(s)}\}_{n=1}^{N}), \tag{8}$$

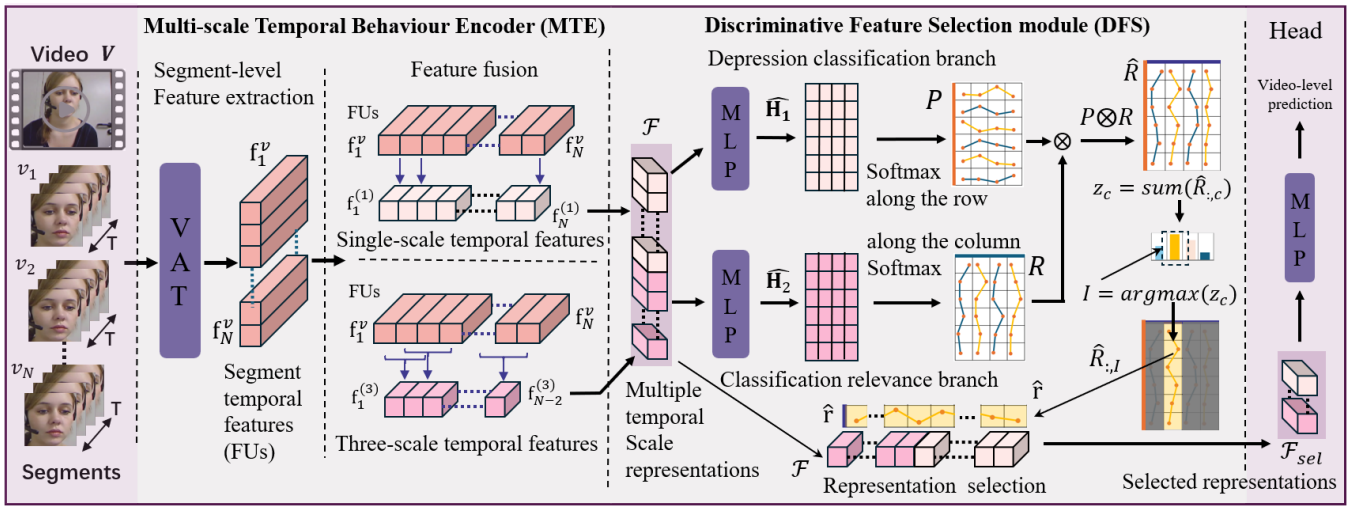

Figure 2: The proposed approach consists of three modules: Multi-scale Temporal Behaviour Encoder (MTE) module, Discriminative Feature Selection (DFS) module and a prediction head. The MTE module takes equal-length segments $\{v_n\}$ as input, extracts segment features $\{\mathbf{f}_\mathbf{n}^\mathbf{v}\}$ and constructs multiple temporal scale facial behaviour representations $\mathcal{F}$. (For the convenience of visualisation temporal scale (s) at 1 and 3 are presented in the figure.) Subsequently, the DFS module digests the $\mathcal{F}$ representations for the selection of depression-related representations, which are denoted as $\mathcal{F}_{\text{sel}}$. Finally, the selected representations are sent to a predictor head for video-level depression assessment.

where $c = 1, \ldots, C$ and C denotes the number of depression categories (e.g., none, mild, modest and severe). $s = 1, \ldots, S$ represents the temporal scales . Accurately analysing video-level depression category from video segments at various temporal scales depends on effectively evaluating the relevance of facial behaviour segments to the video-level depression category. To achieve this, for each depression category c, we explicitly predict relevance scores for all facial behaviour representations across all temporal scales in $\mathcal{F} = \{F^{(s)}\}_{s\in\{1,\ldots,S\}}$ by sending them into a MLP-based classification relevance branch, as:

$$\mathbf{R}_{n,c} = \underset{n}{\text{Softmax}}(\hat{\mathbf{H}}_2) \quad \hat{\mathbf{H}}_2 = \text{MLP}_{\text{rel}}(\{\mathbf{f}_n^{(s)}\}_{n=1}^{N}), \quad (9)$$

The $\mathbf{P} \in \mathbb{R}^{M\times C}$ denotes the matrix of predicted category probabilities from all M facial behaviour representations in $\mathcal{F}$, and $\mathbf{R}_{n,c} \in \mathbb{R}^{M\times C}$ denotes the matrix of predicted relevance scores of each facial behaviour representation in $\mathcal{F}$ to every depression category c. Then the weighted video-level category prediction z_c can be achieved by:

$$\hat{\mathbf{R}} = \mathbf{P} \otimes \mathbf{R} \quad z_c = \sum_{n=1}^{M} \hat{\mathbf{R}}_{n,c} \quad (10)$$

where $\otimes$ denotes element-wise multiply and $c = 1, \ldots, C$. The video-level depression category can be identified as $I = \arg\max_c z_c$.

This updated relevance matrix $\hat{\mathbf{R}}$ can provide a more accurate evaluation of each representation's relevance to the depression category by ensuring the most depression-related representations are also supporting the video-level depression category I. Since the **relevance vector** $\hat{\mathbf{r}} = \hat{R}_{:,I}$ (referred as **segment depression relevance scores** (SDR)) quantifies the relevance of each representation in $\mathcal{F}$ to the video-level depression category prediction, highlighting the segments corresponding to the representations with top-ranked relevance in $\hat{\mathbf{r}}$ can provide segment-level interpretability for video-based ADA, i.e., which segments are most informative to the video-level depression category.

Finally, to mitigate the influence of less relevant or noisy representations, the top K representations, ranked by SDR $\hat{\mathbf{r}} \in \mathbb{R}^M$, are selected and denoted as $\mathcal{F}_p$. These selected representations are passed to a downstream MLP-based prediction head to predict video-level depression scores $y \in \mathbb{R}^K$ individually, where $K = \lfloor p\% \cdot M \rfloor$ and p represents the selection ratio. The final video-level depression score is the average of y.

Training Strategy and Loss Functions

We are the first to formulate the identification of depression-informative facial behaviour segments across multiple temporal scales as a weakly supervised learning task guided by ordinal depression severity categories (none, mild, moderate, and severe) derived from annotated depression scores. This is achieved by propagating the video-level ordinal labels to all segments during training, encouraging the model to focus on temporally localized facial behaviours that align with the overall depression severity. The rationale for identifying depression-relevant facial behaviour segments is grounded in the assumption that, given a ground-truth (GT) depression score (reflecting the severity category) for a video, certain temporal intervals must exhibit facial behaviours that reflect this underlying depression status. However, explicit segment-level annotations are unavailable but only video-level category labels. Inspired by weakly supervised temporal action localization (WSTAL) methods (Ren

et al. 2023; Li, Wang, and Liu 2024), which localize action intervals using only video-level class labels, we reformulate the depression-related interval localization task as a segment-level facial behaviour classification problem. To this end, we aggregate segment-level predictions and align them with the video-level GT depression category, enabling the model to learn depression-relevant segments location by classification under weak supervision. Our method is conceptually simple yet effective, avoiding complex inference mechanisms while achieving accurate and interpretable video-level depression prediction.

Specifically, our training strategy is achieved via two main steps. The first step is to train a category sensitive spatial temporal encoder for depression-related feature extraction. The second step is to further learn how to select depression-informative facial behaviours among multiple temporal scale durations base on the extracted segment-level feature from the encoder.

Training Spatial-Temporal Encoder We first train a VAT model to extract spaito-temporal facial behaviour features from fixed-length video segments (e.g., 32 consecutive frames). During training, the given video V is split into segments $\{v_n\}_{n=1}^{N}$ with equal-length of T frames and an overlap of $T/2$ between the adjacent segments. Each segment is paired with the corresponding ground-truth (GT) video-level depression score y^* $(0-63)$, representing four depression categories (none, mild, modest or severe). The VAT model takes a video segment v_n as input and predicts its depression score $y_n \in \mathbb{R}$ and depression category probability $c_n \in \mathbb{R}^4$. The training loss for depression score regression is the Mean Squared Error (MSE). The spatial-temporal encoder trained with only the depress score regression is referred as **base regression method**. To enhance the model's ability to distinguish depression categories, we incorporate an ordinal classification loss for category prediction and combine it with the MSE loss for VAT model training (referred to as the **ordinal regression method**). For ordinal classification, we treat the four severity categories as three ordered binary tasks. We define indicator targets $t_{n,k} = \mathbb{I}[c_n^* > k]$ for $k \in \{0, 1, 2\}$, which specify whether the GT category exceeds threshold k. The model predicts the corresponding ordered probabilities via a sigmoid function:

$$p_n = \sigma(c_n). \tag{11}$$

The ordinal classification loss is the average binary cross-entropy across all thresholds:

$$\mathcal{L}_{\text{ORD}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=0}^{2}[t_{n,k}\log p_{n,k}+(1-t_{n,k})\log(1-p_{n,k})], \tag{12}$$

where N is the number of segments.

Training Multiple Temporal Scale Representation Selection In the second training step, The MTE and DFS modules are jointly optimized such that the extraction of multiple temporal scale representation is guided by the downstream supervision of video-level depression classification and regression. The MTE module first employs the trained VAT model to extract facial behaviour features from fixed-length non-overlap segments within the given video (referred as facial-behaviour feature units(FUs)). Those FUs are then fused to construct multiple temporal scale representations $\mathcal{F}$ by MTE module, describing facial behaviours at different temporal scales. Subsequently, the DFS module takes each representation $\mathbf{f}_n^{(s)}$ at multiple temporal scale in $\mathcal{F}$ as input and predicts depression category probability $p_{n,c}$ and category relevance scores $r_{c,n}$ individually (Eqa. **??** and **??**). Give the GT depression category c^* as a one-hot vector, the weakly supervised learning (WSL) loss can be formulated as:

$$\mathcal{L}_{\text{WSL}} = -\frac{1}{4}\sum_{i=1}^{4}\left[c_i^*\log(z_i) + (1-c_i^*)\log(1-z_i)\right], \tag{13}$$

$$z_c = \sum_{n=1}^{M}\mathbf{P}_{n,c}\cdot\mathbf{R}_{n,c} \tag{14}$$

where M is the total number of representations in $\mathcal{F}$, and $c = 1, \ldots, 4$. In addition to increase the discriminative ability and suppress the influence of less relevant representations, we select representations evaluated into the top 20% most related ones by Eqa. 5. Those selected representations $\mathcal{F}_{sel}$ are further processed by FC-based head for depression score y and category probability c prediction, where MSE and BCE loss are employed respectively. So the overall training loss for MTE and DFS module can be summarised as:

$$\mathcal{L} = \mathcal{L}_{\text{WSL}} + \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{BCE}} \tag{15}$$

Experiments

Datasets and Implementation details

Dataset: We evaluate our approach on two visual depression assessment benchmark datasets: AVEC 2013 (Valstar et al. 2013) and 2014 (Valstar et al. 2014) as AVEC 2019 (Ringeval et al. 2019) has not released raw videos. Participants in both datasets are labelled with depression scores ranging from 0 to 63 based on the BDI-II. The AVEC 2013 recorded a total of 150 video clips, length in 20 – 50 minutes, from 82 participants. The AVEC 2014 contains two subsets: Northwind and Freeform, each with 150 clips ranging from 6 seconds to 4 minutes. Both datasets were evenly divided into three subsets: training, development and testing.

Metrics: To compare the performance of our approach with others, We follow previous studies (Valstar et al. 2013, 2014; Xu et al. 2024) by measuring root mean square error (RMSE) and mean absolute error (MAE).

Implementation details: We utilized Openface2 (Baltrušaitis, Robinson, and Morency 2016) to obtain face images from video frames and resize them to a resolution of 224×224 pixels. For data augmentation, we applied random horizontal flipping and color jittering. In the first training step, the VAT model was trained for 15 epochs using the Adam optimizer. In the second training step, the MTE and DFS module were jointly trained for 80 epochs using the stochastic gradient descent (SGD) optimizer.

Methods	AVEC 2013		AVEC 2014	
	RMSE↓	MAE↓	RMSE↓	MAE↓
(Meng et al. 2013)	13.61	10.88	10.86	8.86
(Zhu et al. 2017)	9.82	7.58	9.55	7.47
(Al Jazaery and Guo 2018)	9.28	7.37	9.20	7.22
(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
(Song et al. 2022)	8.10	6.16	7.15	5.95
(De Melo, Granger, and Lopez 2020)	7.97	5.69	7.94	6.20
(De Melo, Granger, and Hadid 2020)	7.90	5.98	7.61	5.82
(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
(Niu et al. 2022)	7.42	6.09	7.39	5.87
(Uddin, Joolee, and Sohn 2022)	7.32	5.90	6.98	5.75
(Pan et al. 2023b)	**7.26**	5.97	7.30	5.99
(Niu et al. 2024)	7.49	**5.43**	7.27	5.63
(Xu et al. 2024)	7.57	5.95	7.18	5.86
Ours	7.48	5.77	**6.89**	**5.11**

Table 1: Results achieved for AVEC 2013 and 2014 test sets.

Comparison with Existing Methods

Table 1 compares our proposed approach with existing state-of-the-art methods, where our proposed method achieved competitive performance compared to other depression detection methods on both AVEC datasets. On the AVEC 2013 dataset, our method achieved a RMSE of 7.48, comparable to the performance of state-of-the-art (SOTA) models applied to this dataset. These results highlight the effectiveness of our approach in accurately performing ADA. On the AVEC 2014 dataset, the proposed method achieved SOTA performance with an RMSE of 6.89. This result demonstrates that leveraging video-level depression severity category to guide the selection of the most depression-relevant video segments enhances depression prediction performance. The AVEC 2014 dataset comprises two sub-subsets for the same subjects, Freeform and Northwind. We achieved optimal results by employing a decision-level fusion of video-level predictions from both sub-datasets. Results related to each sub-subset are included in the supplementary materials.

Ablation Study

We conducted comprehensive ablation studies on both datasets. In addition, the analysis of the top $p\%$ relevant representations selection ratio and the segment length T for segment-level facial behaviour feature extraction is presented in Supplementary Material.

Base regression method: As shown in Table 2, The base regression training, trained on video segments only using the MSE loss, achieved RMSE values of 9.15, 9.26, and 9.22, and MAE values of 7.52, 7.40, and 7.71 on the AVEC2013, Freeform, and Northwind datasets, respectively. These results establish the baseline performance of our proposed method, which are used for further evaluation in subsequent experiments. **Ordinal regression method:** To improve the VAT and DFS model's sensitivity to depression category cues, we employed ordinal regression for loss computation, resulting in significant performance improvements. Specifically, on the AVEC 2013 dataset, we observed RMSE and MAE reductions of 9.2% and 11.2%, respectively. On the Freeform dataset, RMSE decreased by 14.3% and MAE by 13.7%, while on the Northwind dataset, RMSE and MAE were reduced by 19.2% and 23.9%, respectively. These accuracy gains highlight the model's improved ability to identify and discriminate facial behaviour segments corresponding to varying category of depression severity, thereby enhancing its predictive accuracy. **Feature selection method:** The proposed the weakly supervised learning method identifies and prioritizes the most depression-related facial behaviour segments for video-level depression prediction, which significantly outperforms the ordinal regression methods that indiscriminately uses all segments. Specifically, RMSE and MAE were reduced by 12.0% and 17.2% on the AVEC 2013 dataset, respectively; by 3.4% and 10.2% on the Freeform dataset and 7.3% and 14.0% on the Northwind dataset.

Dataset	Method	MAE↓	RMSE↓
AVEC2013	Base regression	7.52	9.15
	Ordinal regression	6.76	8.38
	Feature selection	5.77	7.48
Freeform	Base regression	7.40	9.26
	Ordinal regression	6.51	8.10
	Feature selection	5.83	7.64
Northwind	Base regression	7.71	9.22
	Ordinal regression	5.93	7.53
	Feature selection	5.20	7.02

Table 2: Results on AVEC2013 and 2014 (Northwind and Freeform) datasets across various training methods.

Visualization

In this section, we aim to illustrate the advantages of our method in identifying and selecting multiple temporal scale segments for ADA in terms of interpretability, discrimination and effectiveness.

Interpretability: As shown in Figure 3, we plot the segment depression relevance scores (SDR) $\hat{\mathbf{r}}$ for each segment along the temporal axis. Each multiple temporal scale segment is represented by a point, with red, green, and blue colour indicating segments at temporal scales of 32, 64, and 96 frames, respectively. The visualisation result demonstrates that multiple temporal scale segments highlight depression-related facial behaviours across varying durations. The relevance scores fluctuate along the temporal axis, indicating that certain segments are more strongly correlated with video-level depression severity category. These segments highlighted by heatmap underscore the ability of our approach to identify depression-related facial behaviours across various durations.

Discrimination: Figure 4 demonstrates the importance of selecting discriminative segment representations at multiple temporal scales. For each video in the test dataset, we extract the multiple temporal scale representations $\mathcal{F}$ from it and compute the corresponding SDR $\hat{\mathbf{r}}$ for each segment representation, such that the top 20% depression-relevant representations and bottom 20% less relevant rep-

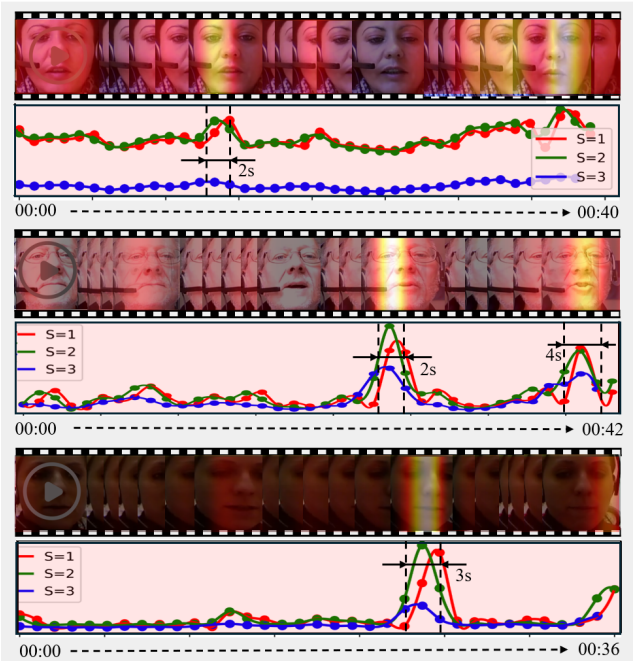

Figure 3: Our approach identifies the most depression-related facial behaviours at multiple temporal scales, shown as the highlighted part in the heatmap for each video, where the darker colour means the less relevant. The figure shows three video's depression-related segments across three temporal scales ($S = 1, 2, 3$), In each visualized video, segments in the same temporal scale are represented by sequential points with the same colour, which denote the relevance scores of segments to the video level depression prediction. The red, green, and blue points indicates segments at temporal scales of 32, 64, and 96 frames respectively.

resentations in each video can be collected respectively to construct a high-relevance representation-set and a low-relevance representation-set. T-SNE is then employed to visualize the representation distributions of these two contrasting sets as scatter plots. As illustrated in the left plot of Fig.4, representations associated with the same depression severity category cluster closely together, while representations associated with different categories are well-separated. This indicates that the representations in the high-relevance dataset are discriminative to their prediction targets. In contrast, as shown in the right plot of Fig.4, representations in the low-relevance set exhibit significant overlap between different severity categories, demonstrating that these representations are not distinguishable. This clear comparison underscores the importance of selecting high-relevant segments for effective video-level depression assessment.

We also evaluate the stability of SDR scores $\mathbf{r}$ for the face videos by computing their cosine similarity among different training rounds. We trained the model 5 times with random initialisations and collected the SDR scores of 3 randomly selected videos after the first training epoch, where the mean pairwise cosine similarities for each video under 5 training times are 0.94, 0.98, and 0.91, indicating the SDR scores are stable across random initialisations.

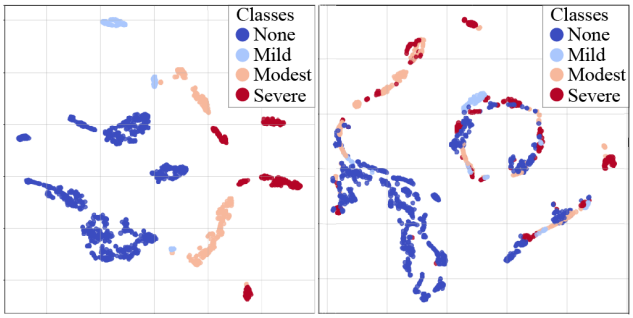

Figure 4: The left plot describes a T-SNE embedding distribution of the top 20% high-relevant representations derived from all videos in the test dataset, while the right describes the bottom 20% low-relevant representations.

Effectiveness: Figure 5 illustrates the deviations of video-level depression predictions from their true labels. The scatter plots demonstrate our method's consistent accuracy in depression detection, evidenced by the alignment of points along the line that best fits.

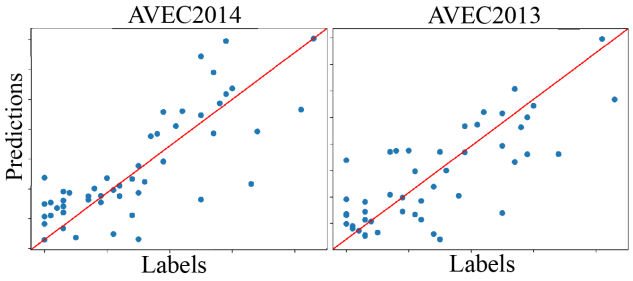

Figure 5: Depression predictions from our approach versus labels on AVEC 2014 and AVEC 2013 datasets.

Conclusion

In this work, we introduced a novel weakly supervised approach that prioritizes variable-length video segments containing depression-related facial behaviours for video-based ADA, while simultaneously improving the interpretability of the ADA process by highlighting behaviour segments associated with depressive symptoms within a target video. However, the highlighted behaviours only reflect the model's internal confidence rather than expert-validated clinical markers. Moreover, the applicability of our approach is currently limited by the availability of suitable datasets, where AVEC 2013/2014 remain the only publicly accessible dataset offering raw facial videos paired with rigours depression labels. Expanding evaluation to larger and clinically verified datasets will be an essential direction for future work to improve the generalizability and the clinical credibility of the approach.

References

Al Jazaery, M.; and Guo, G. 2018. Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing*, 12(1): 262–268.

Baltrušaitis, T.; Robinson, P.; and Morency, L.-P. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, 1–10. IEEE.

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*, 13(3): 1581–1592.

De Melo, W. C.; Granger, E.; and Lopez, M. B. 2020. Encoding temporal information for automatic depression recognition from facial analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1080–1084. IEEE.

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*, 14(1): 578–590.

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

Gahalawat, M.; Rojas, R. F.; Guha, T.; Subramanian, R.; and Goecke, R. 2023. Explainable Depression Detection via Head Motion Patterns. In *Proc. 25th ACM Int. Conf. on Multimodal Interaction (ICMI)*, 261–270.

Girdhar, R.; Carreira, J.; Doersch, C.; and Zisserman, A. 2019. Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 244–253.

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.; Guo, C.; Tiwari, P.; Pandey, H. M.; and Dang, W. 2022. Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence. *International journal of intelligent systems*, 37(12): 10140–10156.

He, L.; Li, Z.; Tiwari, P.; Cao, C.; Xue, J.; Zhu, F.; and Wu, D. 2024a. Depressformer: Leveraging video swin transformer and fine-grained local features for depression scale estimation. *Biomedical Signal Processing and Control*, 96: 106490.

He, L.; Zhao, J.; Zhang, J.; Jiang, J.; Qi, S.; Wang, Z.; and Wu, D. 2024b. LMTformer: facial depression recognition with lightweight multi-scale transformer from videos: LMTformer: facial depression recognition... *Applied Intelligence*, 55(3).

He, L.; Zhao, J.; Zhang, J.; Jiang, J.; Qi, S.; Wang, Z.; and Wu, D. 2025. LMTformer: facial depression recognition with lightweight multi-scale transformer from videos. *Applied Intelligence*, 55(3): 195.

He, L.; Zhao, J.; Zhang, M.; Jiang, J.; Qi, S.; Wang, Z.; and Wu, D. 2024c. LMTformer: Facial Depression Recognition with Lightweight Multi-Scale Transformer from Videos. *Applied Intelligence*, 55(3): 195–210.

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. IEEE.

Li, Z.; Wang, Z.; and Liu, Q. 2024. Weakly supervised temporal action localization with actionness-guided false positive suppression. *Neural Networks*, 175: 106307.

Liao, R.; Song, S.; and Gunes, H. 2024. An open-source benchmark of deep learning models for audio-visual apparent and self-reported personality recognition. *IEEE Transactions on Affective Computing*, 15(3): 1590–1607.

Mahayossanunt, Y.; Nupairoj, N.; Hemrungrojn, S.; and Vateekul, P. 2023. Explainable Depression Detection Based on Facial Expression Using LSTM on Attentional Intermediate Feature Fusion with Label Smoothing. *Sensors*, 23(23): 9402.

Mao, K.; Zhang, W.; Wang, D. B.; Li, A.; Jiao, R.; Zhu, Y.; Wu, B.; Zheng, T.; Qian, L.; Lyu, W.; et al. 2022. Prediction of depression severity based on the prosodic and semantic features with bidirectional LSTM and time distributed CNN. *IEEE transactions on affective computing*, 14(3): 2251–2265.

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.; Li, M.; and Fu, C. 2024. PointTransform Networks for Automatic Depression Level Prediction via Facial Keypoints. *Knowledge-Based Systems*, 297: 111951.

Niu, M.; Li, Y.; Tao, J.; Zhou, X.; and Schuller, B. W. 2024. DepressionMLP: A Multi-Layer Perceptron Architecture for Automatic Depression Level Prediction via Facial Keypoints and Action Units. *IEEE Transactions on Circuits and Systems for Video Technology*.

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*, 14(1): 294–307.

Niu, M.; Zhao, Z.; Tao, J.; Li, Y.; and Schuller, B. W. 2022. Selective element and two orders vectorization networks for automatic depression severity diagnosis via facial changes. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 8065–8077.

Pan, Y.; Shang, Y.; Liu, T.; Shao, Z.; Guo, G.; Ding, H.; and Hu, Q. 2023a. Spatial–Temporal Attention Network for Depression Recognition from Facial Videos. *Expert Systems with Applications*, 237: 121410.

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. 2023b. Integrating deep facial priors into landmarks for privacy preserving multimodal depression recognition. *IEEE Transactions on Affective Computing*.

Ren, H.; Yang, W.; Zhang, T.; and Zhang, Y. 2023. Proposal-based multiple instance learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2394–2404.

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.

Rottenberg, J. 2017. Emotions in depression: What do we really know? *Annual Review of Clinical Psychology*, 13: 241–263.

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.

Stehman, C. R.; Testo, Z.; Gershaw, R. S.; and Kellogg, A. R. 2019. Burnout, drop out, suicide: physician loss in emergency medicine, part I. *Western Journal of Emergency Medicine*, 20(3): 485.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

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*, 13(2): 864–870.

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*.

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.

Wang, R.; Huang, J.; Zhang, J.; Liu, X.; Zhang, X.; Liu, Z.; Zhao, P.; Chen, S.; and Sun, X. 2024. FacialPulse: An efficient RNN-based depression detection via temporal facial landmarks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 311–320.

Wu, Z.; Zhou, L.; Li, S.; Fu, C.; Lu, J.; Han, J.; Zhang, Y.; Zhao, Z.; and Song, S. 2025. DepMGNN: Matrixial Graph Neural Network for Video-based Automatic Depression Assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1610–1619.

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*.

Yan, K.; Miao, S.; Jin, X.; Mu, Y.; Zheng, H.; Tian, Y.; Wang, P.; Yu, Q.; and Hu, D. 2024. TCEDN: A Lightweight Time-Context Enhanced Depression Detection Network. *Life*, 14(10): 1313.

Yang, M.; Shang, Y.; Liu, J.; Shao, Z.; Liu, T.; Ding, H.; and Li, H. 2024. LMS-VDR: Integrating Landmarks into Multi-scale Hybrid Net for Video-Based Depression Recognition. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 299–312. Springer.

Zhang, J.; Huen, J. M. Y.; Lew, B.; Chistopolskaya, K.; Talib, M. A.; Siau, C. S.; and Leung, A. N. M. 2020. Depression, anxiety, and stress as a function of psychological strains: Towards an etiological theory of mood disorders and psychopathologies. *Journal of affective disorders*, 271: 279–285.

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*, 15(3): 1078–1089.

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*, 11(3): 542–552.

Zhu, Y.; Shang, Y.; Shao, Z.; and Guo, G. 2017. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing*, 9(4): 578–584.