Learning from Label Relationships in Human Affect

Niki Maria Foteinopoulou
Queen Mary University of London
London, United Kingdom
n.m.foteinopoulou@qmul.ac.uk

Ioannis Patras
Queen Mary University of London
London, United Kingdom
i.patras@qmul.ac.uk

ABSTRACT

Human affect and mental state estimation in an automated manner, face a number of difficulties, including learning from labels with poor or no temporal resolution, learning from few datasets with little data (often due to confidentiality constraints) and, (very) long, in-the-wild videos. For these reasons, deep learning methodologies tend to overfit, that is, arrive at latent representations with poor generalisation performance on the final regression task. To overcome this, in this work, we introduce two complementary contributions. First, we introduce a novel relational loss for multilabel regression and ordinal problems that regularises learning and leads to better generalisation. The proposed loss uses label vector inter-relational information to learn better latent representations by aligning batch label distances to the distances in the latent feature space. Second, we utilise a two-stage attention architecture that estimates a target for each clip by using features from the neighbouring clips as temporal context. We evaluate the proposed methodology on both continuous affect and schizophrenia severity estimation problems, as there are methodological and contextual parallels between the two. Experimental results demonstrate that the proposed methodology outperforms the baselines that are trained using the supervised regression loss, as well as pre-training the network architecture with an unsupervised contrastive loss. In the domain of schizophrenia, the proposed methodology outperforms previous state-of-the-art by a large margin, achieving a PCC of up to 78% performance close to that of human experts (85%) and much higher than previous works (uplift of up to 40%). In the case of affect recognition, we outperform previous vision-based methods in terms of CCC on both the OMG and the AMIGOS datasets. Specifically for AMIGOS, we outperform previous SoTA CCC for both arousal and valence by 9% and 13% respectively, and in the OMG dataset we outperform previous vision works by up to 5% for both arousal and valence.

KEYWORDS

continuous affect estimation, multilabel, representation learning

1 INTRODUCTION

Understanding human affect and mental state is an active research area with multiple potential applications spanning fields such as education [49], healthcare [41], and entertainment [32, 40]. For example, by understanding human emotion, the user experience can be enhanced and healthcare professionals can more effectively monitor the patients' emotional state. These problems can be treated either as a classification, using the basic human emotions [14] or by utilising continuous labels along the Arousal-Valence axes [39]. Similarly, in the domain of mental illness, several scales have been used by healthcare professionals to assess the severity of the symptoms, thus treating symptoms as a spectrum [2].

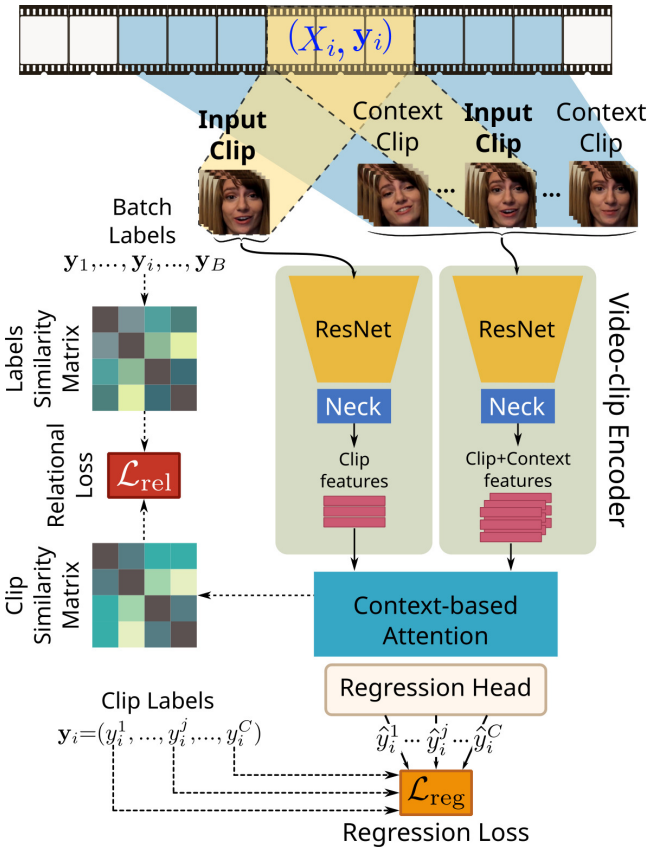


Figure 1: Overview of the proposed framework.

Regardless of which of the above labelling approaches is adopted, certain issues render the problem of human affect and mental state estimation challenging. Specifically, in-the-wild datasets tend to include long videos with low or no temporal label resolution – i.e., a set of labels describes the entire video. This typically occurs as affect and mental health symptom labels refer to abstract behaviour that is not easily captured and is not always objectively defined. The length of the video poses a major difficulty for Machine Learning methods, due to GPU memory constraints. To address this issue, two main approaches are employed in the literature, namely, a) estimating sub-segments of the long videos [8, 31] and b) precomputing features[5, 35, 50, 51]. For example, in MIMAMO [12] and the work of Peng *et al.* [36] a small number of frames is sampled from each clip. However, this disregards information from the remaining video and the clip context. Moreover, as affect and mental state descriptions often refer to a larger context, short clips might

not be representative samples. Similarly, estimating per-frame predictions [33] disregards clip information and is also suffering from the lack of temporal information. Previous state-of-the-art works in symptom severity estimation [4], used statistical representations, such as Gaussian Mixture Models, on a set of per-frame extracted features. However, this approach does not learn from the temporal relationships of frame features. It also does not allow for end-to-end training, therefore does not allow for feature optimisation on the specific task. In order to exploit contextual information and improve clip-level features, Wu et al. [48] proposed the use of Long-Term Feature Banks for the problem of action recognition in videos. However, Long-Term Feature Banks [48] rely on a pre-computed set of features for the context, that does not improve in quality during training. By contrast, in this work, we build upon [48] and use a context feature extractor that updates context features at each iteration, allowing for dynamically computing context features of random clips sampled from a longer video in an end-to-end manner, leading to much shorter training times.

Publicly available datasets for affect and mental health analysis are typically small, which often results in overfitting problems during training. As such, methods that lead to better representations with a small number of samples are paramount to the success of the final regression task. However, several recent works [6, 11, 20, 26] require pre-training (whether supervised or unsupervised) with very large datasets to achieve better representations before fine-tuning on the final task. In continuous affect estimation, Kim *et al.* [27] binarised labels and used an adversarial loss on the latent feature space, however this approach ignores the continuous nature of Arousal/Valence dimensions.

In order to both alleviate the challenges due to long video input and to improve the feature representations so as to address the multi-label regression problems that arise in the domain of affect and mental health analysis, in this work we propose a) a novel attention-based video-clip encoder that builds upon [48] and utilises the temporal dimension of the input clips and arrives at clip-level predictions that benefit from context clip information, and b) a novel relational regression loss function that aligns the distances in the latent clip-level representations/features to the distances of the labels of the clips in question. An overview of the proposed framework is shown in Fig. 1. Specifically, we propose to jointly train two network branches: a) one that uses the proposed video-clip encoder to extract clip-level features from the input video clips and a set of temporally neighbouring clips, which subsequently feed a regression head in order to infer the desired values and calculate the regression loss, and b) one that uses the proposed video-clip encoder to extract clip-level features from the input video clips, which subsequently feeds the regression head and are further used to construct the intra-batch similarity matrix for calculating the proposed relational loss. To the best of our knowledge, this is the first work that uses label relationships to improve feature representation learning. The proposed regression head employs an attention-based mechanism for fusing clip-level and context features and regressing to the desired continuous values. The main contributions of this work can be summarised as follows:

- We build on [48] and propose a two-stage attention architecture that uses features from the clips' neighbourhood to introduce context information in the feature extraction. The architecture is novel in the domain of affect and mental state analysis and, unlike [48], it does not train a separate model to compute context features, but rather updates its weights during training – this leads to smaller training times.
- We introduce a novel loss, named relational regression loss, that aims at learning from the label relationships of the samples during training. This loss is using the distance between label vectors to learn intra-batch latent representation similarities in a supervised manner. We show in the ablation studies that the improved latent representations obtained with the addition of the relational loss lead to improved regression output, without the use of large datasets.
- We show that the methodology achieves results comparable to the state-of-the-art. Specifically, for symptom severity estimation of schizophrenia, our methodology outperforms the previous state of the art on all scales and symptoms tested and achieves a Pearson's Correlation Coefficient similar to that of human experts.

2 RELATED WORK

In this Section, we review previous works on continuous affect estimation and mental health assessment and focus on representation learning and temporal context exploitation in large videos.

2.1 Learning Representations and Label Relations

In human affect problems and even more in mental state estimation, learning features representative of the behaviour rather than other entangled factors (eg. identity) is paramount to the reliability of the final estimate. A number of works have addressed the issue of representation learning, with more recent developments in contrastive methodologies [10, 11], whether evaluating results on static image data or video datasets [38]. These self-supervised methodologies learn latent representations by teaching the architecture which data points are similar. By extending the idea of comparing samples, supervised contrastive frameworks propose that images [26] or videos [20] from the same class are treated as similar, which results in embeddings from the same class being more closely aligned. However, these works are trained on very large datasets which are not typically available in affective and mental health problems, and have only been evaluated on classification problems. Kim *et al.* [27] implement an adversarial loss to learn better representations for continuous affect, however, Arousal/Valence values are binarised for the adversarial task. In our work, we explore the idea of learning representations by comparing sample similarities in a supervised approach, however we implement a non-binary approach which is more suitable for multi-label regression problems.

Several problems/datasets in the field of continuous affect and mental health have multiple labels, in order to describe various affective attributes/ psychological symptoms. Treating each label independently [12] ignores their potential correlations as well as increases training times significantly with each additional label. Several works investigate multi-label recognition problems using graph learning approaches to model label correlations and co-occurrences [30, 47]. However, such approaches do not learn

from label similarities *between* samples and do not project these similarities to the latent representation space. In contrast, our work uses information from the inter-sample label similarities to learn better latent representations.

2.2 Addressing large sequences

The exploration of methods tailored to long-range video understanding, is vital for human affect and mental state estimation, as long videos are typically more representative of real-life settings. Moreover, long temporal relationships intuitively should contribute to more accurate estimates of human affect and mental states. To address long video sequences, previous works have used a number of strategies. One such method is to pre-compute features [4, 29]; this however does not allow for end-to-end training and makes augmentation techniques more complicated (if feasible at all). Another strategy to address long video sequences is by using contextual features either in the form of intra-sample relations [52] or by exploring feature banks [48]. Both of these approaches utilise relations between the short term actions, which is the temporal context of a clip. However, these methods have been evaluated on action recognition problems and have not been implemented in affect. Moreover, while action problems benefit from from long-term context, they still have a much lower label resolution. Finally, in [48] context features need to be pre-computed on pre-defined clips, therefore feature quality does not improve with further training.

In our work, we build on the concept of exploiting contextual features, however differently from [48] we use context features to improve clip-level prediction with end-to-end training. We also do not operate on pre-defined clips, but rather dynamically compute features from them – with this approach as training progresses, the network learns from better context features.

2.3 Affect and Mental Health

As several mental health illnesses and disorders have non-verbal behaviour symptoms, understanding patients' affect is important in diagnosis and severity estimation. Depression is a mood disorder that has an impact on patients' affective state; similarly, a number of schizophrenia negative symptoms refer to patients' affect and expressions, therefore there are important semantic parallels between continuous affect estimation and mental health assessment, so we examine them in parallel. More work has been performed on estimating depression severity than symptoms of schizophrenia, as there are no publicly available datasets for the latter.

A number of previous works use a bag of words approach for gestures and facial expressions [22, 23] or use statistical representations of pre-computed features [5, 16, 43]. Some of the symptoms have a quantitative measure (e.g., reduced gestures), therefore, intuitively, methodologies that implicitly measure quantity of features have achieved state-of-the-art results in these previous works. However, such methods disregard the temporal relationship of features. Similarly in [33], by making per-frame predictions for depression estimation, the temporal dimension is not taken into consideration. More recently, in [21] the temporal dimension is taken into consideration as the work leverages audio and text modalities, however, no modality for vision is implemented. In contrast, our work proposes a transformer-based architecture to learn from the temporal dimension. We implement this on RGB frames cropped to the subjects' faces, rather than a set of pre-computed features such as Facial Action Units [14].

3 PROPOSED METHOD

An overview of the proposed framework for the problem of multi-label regression from a sequence of clips is given in Fig. 2. In a nutshell, the proposed architecture consists of two branches with shared weights, that incorporate two main components: a) a video-clip encoder employing a convolutional backbone network for frame-level feature extraction and b) a Transformer-based network leveraging the temporal relationships of the spatial features for clip-level feature extraction (Sect. 3.1). The clip and context features produced by the aforementioned branches are passed to a context-based attention block (Sect. 3.2) and a regression head (Sect. 3.3). The proposed method uses the context-based attention block to incorporate features from the two branches before passing them to the regression head, as shown in Fig. 2. The bottom branch uses the proposed video-clip encoder to extract clip-level features from the input video clips, which subsequently feed the context-based attention block and are further used to construct the intra-batch similarity matrix for calculating the proposed relational loss (Sect. 3.4). The goal of the proposed relational loss, as an additional auxiliary task to the main regression, is to obtain a more discriminative set of latent clip-level features, by aligning the label distances in the mini-batch to the latent feature distances. Finally, the upper branch uses the proposed video-clip encoder to extract clip-level features from the input video clips and a set of context clips from each of the input clips, which subsequently feed the regression head in order to infer the desired values and calculate the regression loss.

3.1 Video-clip encoder

Let $\mathcal{X}$ be a batch of labelled clips designed so as it contains consecutive clips taken from different video sequences; i.e., $\mathcal{X} = \{(X_i, \mathbf{y}_i)\}_{i=1}^{B}$, where $X_i \in \mathbb{R}^{T\times H\times W\times 3}$ denotes the i-th clip in the mini-batch, T denotes its duration in frames, H, W denote the frame height and width, $\mathbf{y}_i = (y_1, \ldots, y_C) \in \mathbb{R}^C$ denotes the corresponding ground truth label vector with continuous annotation for C classes, and B denotes the mini-batch size.

Given an input clip X_i, the proposed video-clip encoder extracts frame-level features by feeding them to a backbone convolutional network (e.g., a ResNet [19]), which subsequently feeds a Transformer-based network for extracting clip-level features, leveraging this way the temporal relationships of the calculated spatial features. In the proposed framework, we use the above video-clip encoder in both branches as shown in Fig. 2 – i.e., for calculating the clip-level features $\mathbf{z}_i^0 \in \mathbb{R}^D$ for the input clips X_i, $i = 1, \ldots, B$ (bottom branch) and for calculating clip-level features $Z_i = \left(\mathbf{z}_i^{-K}, \ldots, \mathbf{z}_i^0, \ldots \mathbf{z}_i^K\right) \in \mathbb{R}^{(2K+1)\times D}$ from each X_i along with a number K of context clips before and after it (upper branch).

3.2 Context-based Attention

As discussed above, for any given clip X_i and $2K$ context clips around it, the proposed video-clip encoders extract the clip-level

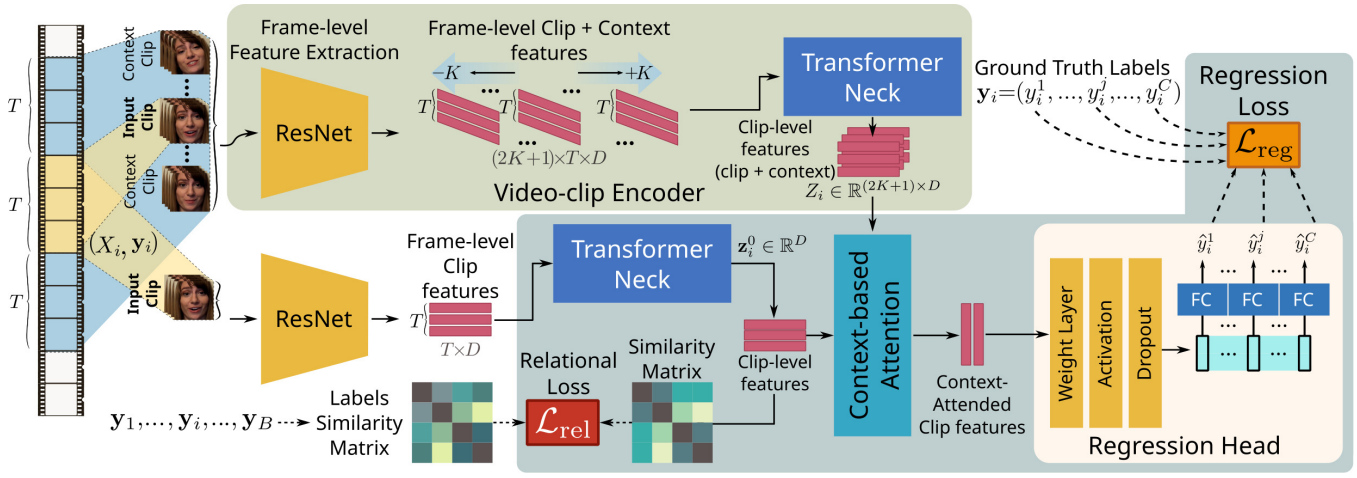


Figure 2: Overview of the proposed framework: (a) The *bottom branch* uses the proposed video-clip encoder (comprising of a ResNet frame-level and a Transformer clip-level feature extractors) to extract clip-level features from the input video clips, which subsequently feed the context-based attention block and are further used to construct the intra-batch similarity matrix for calculating the proposed relational loss. (b) The *upper branch* uses the proposed video-clip encoder to extract clip-level features from the input video clips and a set of context clips from each of the input clips, which subsequently feed the context-based attention block in order to infer the desired values and calculate the regression loss. The context-based attention block, fuses clip-level and context features and passes the context attended clip features to the regression head that estimates the desired continuous values. Error is back-propagated only through the shaded region of the bottom branch.

features $\mathbf{z}_i^0 \in \mathbb{R}^D$ (corresponding to the input clip X_i alone) and $Z_i = \left(\mathbf{z}_i^{-K}, \ldots, \mathbf{z}_i^0, \ldots \mathbf{z}_i^K\right) \in \mathbb{R}^{(2K+1)\times D}$ (corresponding to the input clip X_i and K clips before and K clips after it). These features are then fed to the regression head (Fig. 2), where they are first passed through an attention module before being concatenated. The resulting context-attended clip features are passed to the regression head for the final regression task.

3.3 Multi-label regression head

The context-attended clip features obtained through staged attention as explained in the previous sections, is passed through an MLP regression head that predicts the regression values $\hat{\mathbf{y}}_i = (\hat{y}_i^1, \ldots, \hat{y}_i^C)$, $i = 1, \ldots, C$. Finally, we calculate the regression loss $\mathcal{L}_{\text{reg}}$ by either using the Root Mean Square Error (RMSE) or the Concordance Correlation Coefficient (CCC), depending on the task at hand, as we will discuss in Sect. 4.

3.4 Relational loss

At each training iteration, after having calculated (as discussed in Sect. 3.1) the clip-level features for the clips in a mini-batch, i.e., $\mathbf{z}_i^0 \in \mathbb{R}^D$, $i = 1, \ldots, B$, we calculate the proposed relational loss as follows:

$$\mathcal{L}_{\text{rel}} = \sqrt{\frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \left(\hat{M}_{i,j} - M_{i,j}\right)^2}, \tag{1}$$

where $\hat{M} \in \mathbb{R}^{B\times B}$ denotes the cosine similarity matrix calculated on the clip-level features, whose (i, j)-th element is given as

$$\hat{M}_{i,j} = \frac{\mathbf{z}_i^0 \cdot \mathbf{z}_j^0}{\|\mathbf{z}_i^0\| \|\mathbf{z}_j^0\|},$$

and $M \in \mathbb{R}^{B\times B}$ denotes the cosine similarity matrix calculated on the ground truth labels, whose (i, j)-th element is given as

$$M_{i,j} = \frac{\mathbf{y}_i \cdot \mathbf{y}_j}{\|\mathbf{y}_i\| \|\mathbf{y}_j\|}.$$

It is worth noting that, for the calculation of the proposed relational loss, we use the clip-level features from the given clips without using any context clips, in contrast to the regression loss where additional context clips are being used, as discussed in Sect. 3.3. The total loss is then calculated as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg}} + \lambda \mathcal{L}_{\text{rel}}$, where λ is a weighting hyper-parameter which we discuss in Sect. 4.

3.5 Implementation details

3.5.1 Backbone frame-level feature extractor. We use a standard ResNet50 [19] pre-trained on VGGFace2 [7] and fine-tuned on FER2013 [18] as described in [1]. The classification layer of the pre-trained network was replaced with a fully connected (FC) layer that was fine-tuned for our task during the training of the network, followed by a ReLU [17] activation. The adopted backbone network receives an input of shape $H \times W \times 3$, where H, W are the height and width of the input frame, respectively, and are set to 224 pixels, and outputs a feature vector with 2048 dimensions for each frame. The per-frame feature vectors are stacked to a matrix of size $T \times 2048$ for each clip, where T is the number of frames of each input clip.

3.5.2 Transformer neck clip-level feature extractor. A transformer encoder architecture is employed to learn from the temporal relationships of the spatial feature vectors calculated by the convolutional frame-level feature extractor. The $T \times 2048$ features are positionally encoded and fed forward to a Transformer Encoder [46]. An element-wise addition is performed between the transformer encoder output and the frame-level features, followed by an average pooling operation along the temporal dimension, resulting in a D-dimensional clip-level representation, where $D = 2048$.

3.5.3 Context-base Attention. For each input clip X_i, the regression head takes as input both the clip-level features $\mathbf{z}_i^0 \in \mathbb{R}^D$ and the stacked context features $Z_i = \left(\mathbf{z}_i^{-K}, \ldots, \mathbf{z}_i^0, \ldots \mathbf{z}_i^K\right) \in \mathbb{R}^{(2K+1)\times D}$ (Sect. 3.1). A modified non-local block [48] is then used as an attention operation, where clip-level features $\mathbf{z}_i^0$ are used as the query values to attend to features in Z_i, which are used as keys and values. The output context attention vector is concatenated with the clip-level features, resulting in a $2 \times D$ dimensional vector.

3.5.4 Regression head. The penultimate feature vector is obtained by passing the context-attended feature vector through an FC layer followed by a ReLU activation and a dropout layer.

Finally, in order to obtain the final regression predictions, we split the aforementioned penultimate feature vector into C subsets and attach an FC layer to each subset to obtain the final regression predictions. In the case of continuous affect estimation, we set $C = 2$ (i.e., for Arousal/Valence estimation), while for the schizophrenia symptom severity estimation, we set C accordingly to the number of symptoms provided by the scale at hand. Specifically, the CAINS-EXP scale has 4 symptoms in total therefore we set $C = 4$. The PANSS Negative scale has 7 symptoms in total, however we select 3 for comparison with previous works [4, 45]. As the PANSS-NEG scale includes a number of symptoms we do not consider, we add an additional subset in the penultimate feature vector so that $C = 4$, which is only considered in the total score estimation. We note that, in the case of symptom severity estimation, additionally to each individual symptom prediction, we predict a total score (by using an additional FC layer) using the entire aforementioned penultimate feature vector. This is in contrast to [4], where the total score is estimated using the individual symptom scores.

4 EXPERIMENTAL SETUP

4.1 Datasets

NESS: The dataset was originally collected to study the effect of group body psychotherapy on negative symptoms of schizophrenia [37]. The participants in this study were recruited from mental health services from different parts of the UK. In total 275 participants were interviewed at three different stages of the study: a) a baseline, b) at the end of the treatment, and c) after six months. Each clinical interview recording is between 40-120 minutes long and is performed in-the-wild, reflecting this way the conditions of real-life clinical interviews. Each interview is assessed in terms of two symptom scales, namely, PANSS [24] and CAINS [15]. Out of the total 275 patients, 110 accepted to be recorded at baseline, 93 at end of treatment, and 69 in the six months follow up. The videos in the dataset were recorded at various resolutions and frames per second, however we standardised the resolution to 1920×1080 and fps to 25 frames/s for all videos and we discarded videos where a face was not detected on more than 10% of the frames. Training and evaluation were performed on videos recorded at baseline, for a fair comparison with works in the literature, i.e., 113 videos for 69 patients. All results reported on this dataset are based on a leave-one-patient-out cross-validation scheme, where videos were down-sampled to 3 fps. The values for "Total Negative" and "EXP - Total" in the PANSS and the CAINS scales, respectively, were scaled during training to match the range of individual symptoms (i.e., 1-7 for PANSS and 0-4 for CAINS).

Table 1: Performance (CCC) of the proposed method against baseline and other uni-modal architectures (OMG).

	Arousal	Valence
Proposed	<u>0.26</u>	**0.48**
Proposed w/o K	**0.29**	<u>0.46</u>
Proposed w/o K w/o $\mathcal{L}_{rel}$	0.24	0.44
Proposed w/ $\mathcal{L}_{cont.}$	0.15	0.32
Peng *et al.* [36]	0.24	0.43
Kollias and Zafeiriou [28]	0.13	0.40

OMG: The "OMG-Emotion Dataset" [3] consists of in-the-wild videos of recorded monologues and acting auditions, collected from YouTube. Multiple annotators separated each clip into utterances and assigned labels for Arousal in the $[0, 1]$ scale and Valence in the $[-1, 1]$ scale. The dataset originally consisted of training, validation, and test sets with a total of 7371 utterances. As a number of videos have been removed since the publication of the dataset, we trained on 2071 and evaluated on 1663 utterances. We also scaled Arousal to $[-1, 1]$ to match the range of Valence during training and inference.

AMIGOS: The AMIGOS dataset [34] consists of audio-visual and physiological responses of participants (either alone or in a group) to a video stimulus. In this work, we used the responses of individuals; i.e., where 40 participants watched 16 short videos and 4 long ones. The former were defined as videos of 50-150 seconds. The responses were broken down to 20-second intervals and annotated by three annotators for *Arousal* and *Valence* on a $[-1, 1]$ scale. We extracted the frames from the video (6 frames/s) and calculated the average score of the three annotators as the ground truth during training for the video segment. We trained the network following a leave-one-subject-out cross validation scheme. At each fold we randomly selected a subset of the training data, corresponding to 20% of samples. This is to show how the relational loss can achieve state-of-the-art results using a much smaller number of samples than conventional supervised methodologies.

4.2 Augmentation

During training, we applied data augmentation to the spatial dimensions of all datasets. Specifically, we randomly changed the contrast, the saturation, and the hue of frames with a factor of 0.2 and we applied random horizontal flipping and random rotations (with a range of 30°). The same set of transformations was applied to all frames within a clip. Moreover, as clips with temporal length T were selected from a larger video, we considered the clipping

Table 2: Effect of number of frames T in terms of CCC (OMG).

	Arousal	Valence	Mean
$T = 8$	0.25	0.41	0.33
$T = 16$	**0.26**	**0.49**	**0.375**
$T = 32$	0.19	0.40	0.295

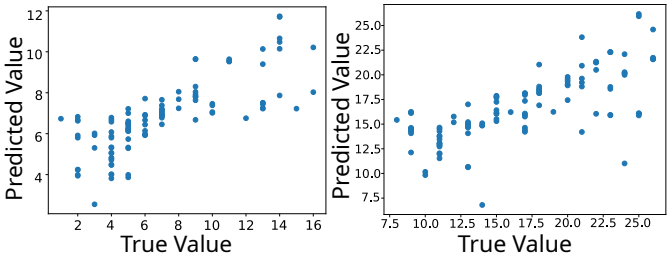
along the temporal dimension as an augmentation approach. More specifically, from the video sequence, we selected a random initial frame and selected T consecutive frames to form a clip. Similarly, the context clips were defined as clips with T number of frames that were positioned before and after the current clip in the video sequence. If the initial frame selected did not allow us to define a complete clip, we looped the video. The number of frames T was set to 32 for the experiments conducted on the NESS, and to 16 for the experiments conducted on the OMG and the AMIGOS datasets.

4.3 Training

During training, the hyperparameter λ that scales the relational loss was empirically set to 2 for experiments conducted on the NESS and the AMIGOS datasets, and to 1 for experiments conducted on the OMG dataset. During testing, the clips were generated by a sliding window over the video sequence, resulting in non-overlapping clips; the average prediction of all clips in the video was calculated to estimate the final predicted label vector. The network was trained in an end-to-end manner with a batch size of 4, 8, and 16 for the NESS, the OMG, and the AMIGOS datasets, respectively, keeping the pre-trained weights of the ResNet-50 backbone frozen. We used an Adam optimizer with an initial learning rate of 10^{-4}, multiplied by 0.1 every 5 epochs, and weight decay $5 \cdot 10^{-3}$. The hyperparameter K that controls the context window size was set to 2 for the experiments on the NESS and to 1 for the experiments on the OMG and the AMIGOS datasets. The network incorporated an RMSE loss during training for the experiments conducted on the NESS and (1-CCC) for the experiments on the OMG and AMIGOS datasets, as proposed by previous works in continuous affect [12, 42].

4.4 Architecture Complexity

The proposed architecture has 90M trainable parameters, distributed as 4M in the backbone, 52M in the transformer neck and 33M in the context aware attention and regression head. We note that, even though the architecture is using two branches (one for clip level features and one for context features), the two branches share weights which significantly reduces the number of parameters. We also note, that similarly to other state-of-the-art methods [12, 28], we use a ResNet50 as our backbone network, but in contrast to them that employ an RNN architecture to explore the temporal relationships, we instead use a Transformer Encoder module. As shown in [46], the self-attention layers of the Transformer are both faster and less complex than recurrent layers (RNN) when the sequence length is shorter than the feature dimensionality, which is the case in the current architecture, hence, the proposed method is more efficient than RNN-based two-stream methods. We report that the inference time is on average at 28.6ms (±2ms) for a clip prediction.



(a) CAINS: EXP – Total | PCC: 0.77 (b) PANSS: NEG – Total | PCC: 0.71

Figure 3: Scaled "Total Score" estimations of the proposed method on NESS using (a) CAINS and (b) PANSS scales.

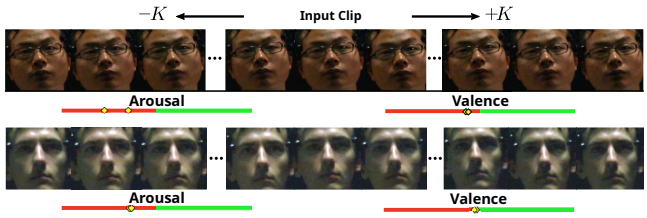


Figure 4: Examples of input clips, their context, and proposed method output on the AMIGOS [34] dataset: In the top row our method predicted A:−0.22**, V:**−0.12 **(ground truth: A:**−0.42**, V:**−0.12**), and in the bottom row A:**−0.29**, V:**−0.03 **(ground truth: A:**−0.29**, V:**−0.04**).**

5 RESULTS AND DISCUSSION

In this section we present the experimental evaluation of the proposed framework. We begin with our ablation study in Sect. 5.1, in order to demonstrate the effectiveness of our method with respect to various design options. Then, in Sect. 5.2, we present comparisons with state-of-the-art methods, where we show that the proposed method achieves results comparable to the state-of-the-art – specifically, for symptom severity estimation of schizophrenia, our method outperforms the previous state-of-the-art on all scales and symptoms tested and achieves a Pearson's Correlation Coefficient similar to that of human experts.

5.1 Ablation study

In order to examine the effect of number of frames T in the overall method, we train the proposed methodology for $T = 8, 16, 32$ on the OMG and NESS datasets. The results of the ablation on T for the OMG dataset is shown on Table 2; we observe that the highest *CCC* for both arousal and valence is achieved when $T = 16$, closely followed by $T = 8$. The effect of T on the PANSS-NEG scale are shown on Table 3; we note that model performance is overall benefited by a larger T, with the exception of symptom N6, which is consistent with the symptom definition (i.e., Lack of Spontaneity and Flow in conversation, which is expected to be short-termed).

In order to investigate the effectiveness of the components of the proposed framework, we conducted an ablation study where we gradually excluded the incorporation of contextual clips and the proposed relational loss. For doing so, we trained a baseline

Table 3: Effect of T on the PANSS-NEG symptom scale.

	N3: Poor Rapport			N6: Lack of Spontaneity			N1: Blunted Affect			Total Negative		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
$T = 8$	0.97	1.31	0.67	**0.61**	**0.80**	**0.66**	**0.62**	0.88	0.55	3.27	4.14	0.64
$T = 16$	0.98	1.29	0.70	0.68	0.86	0.62	0.68	0.93	0.45	3.59	4.54	0.54
$T = 32$	**0.87**	**1.16**	**0.78**	0.74	0.95	0.47	0.64	**0.87**	**0.56**	**2.80**	**3.78**	**0.71**

Table 4: Ablation study on the PANSS-NEG symptom scale.

	N3: Poor Rapport			N6: Lack of Spontaneity			N1: Blunted Affect			Total Negative		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Proposed	**0.87**	**1.16**	**0.78**	**0.74**	**0.95**	**0.47**	**0.64**	**0.87**	**0.56**	**2.80**	**3.78**	**0.71**
Proposed w/o $\mathcal{L}_{rel}$	1.09	1.41	0.66	0.82	0.98	0.44	0.74	0.95	0.39	3.51	4.34	0.66
Proposed w/o $\mathcal{L}_{rel}$ w/o K	1.25	1.57	0.41	0.85	1.01	0.36	0.75	0.97	0.36	3.70	3.62	0.58
Proposed w/ $\mathcal{L}_{cont.}$	1.09	1.53	0.41	0.88	1.05	0.19	0.77	1.01	0.35	3.56	4.48	0.55

Table 5: Ablation study on the CAINS-EXP symptom scale.

	Facial Expression			Vocal Expression			Expressive Gestures			Quantity of Speech			EXP-Total Score		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Proposed	**0.56**	**0.72**	**0.75**	**0.65**	**0.89**	**0.71**	**0.71**	**0.89**	**0.76**	**0.60**	**0.82**	**0.54**	**1.88**	**2.6**	**0.77**
Proposed w/o $\mathcal{L}_{rel}$	0.59	0.78	0.63	0.75	0.98	0.60	0.72	0.96	0.59	0.62	0.85	0.51	2.12	2.94	0.71
Proposed w/o $\mathcal{L}_{rel}$ w/o K	1.06	1.33	0.64	1.06	1.36	0.59	1.14	1.37	0.62	0.77	1.02	0.44	3.76	4.72	0.48
Proposed w/ $\mathcal{L}_{cont.}$	1.12	1.37	0.45	1.06	1.35	0.54	1.19	1.49	0.41	0.84	1.11	0.26	3.87	4.80	0.41

network without context features and trained only on the standard regression loss (i.e., without the proposed relational loss), which we denote as "w/o K w/o $\mathcal{L}_{rel}$". We also trained a version of the network including the context branch without the relational loss, which we denote as "w/o $\mathcal{L}_{rel}$". We finally conducted an experiment using an unsupervised contrastive pre-training, which we denote as "$\mathcal{L}_{cont.}$". In this scenario, we firstly pre-trained the clip-level feature extraction backbone in an unsupervised contrastive manner, and then we trained the regression head on top of the frozen backbone, using the regression loss. For the unsupervised contrastive loss, we sampled 2 clips from the same video as positive samples and considered samples from other videos as negatives.

The analysis results on the NESS [37] dataset are shown in Tables 4, 5 for the PANSS and CAINS scales respectively. We see that the proposed network under the contrastive pre-training scenario, has a similar performance to experiments where we trained with only the regression loss (shown as "w/o $\mathcal{L}_{rel}$") in terms of MAE/RMSE, however in terms of PCC the non-contrastive network still outperforms the contrastive methodology by a large margin. We attribute this to the size of the dataset that was required to learn discriminative features, as other unsupervised methodologies for representation learning [10, 11, 38] trained on very large datasets such as ImageNet [13] and Kinetics [9, 25]. Furthermore, the proposed relational clearly leads to a large improvement to the overall regression task, against the baseline and the unsupervised contrastive loss using a small number of training samples. Contextual features also improved the overall regression performance particularly for the MAE/RMSE metrics, with a more noticeable improvement in the Total Scores of the two scales.

The results of our ablation study on the OMG dataset [3] are presented in Table 1. Comparing the proposed methodology against its baseline (i.e., "w/o K w/o $\mathcal{L}_{rel}$"), we observe that the proposed relational loss improves the performance of the regression measured in terms of CCC, for both Arousal and Valence. Further incorporating the contextual features improved the CCC score for Valence, but lowered slightly the CCC for Arousal. However, compared to other works submitted to the challenge [28, 36], the proposed network and specifically the use of the novel relational loss, shows a clear improvement in terms of CCC for both Arousal and Valence. We also observe a clear advantage of the proposed method compared to the architecture pre-trained with contrastive loss, which appears to over-fit and it may be encouraging the network to learn features of the subjects' identities rather than affective and mental states, due to the nature of the problem and database size.

5.2 Comparison to state-of-the-art

In this section we present the results of the proposed method against state-of-the-art methods. The results for the NESS dataset [37] against previous works are shown in Tables 6 and 7, for PANSS and CAINS scales respectively. We can see that the proposed methodology outperforms previous works across all the evaluated symptoms and scales by a large margin, particularly for PCC, achieving state-of-the-art results. Since the NESS dataset has been annotated by different healthcare professionals, we can compare the PCC achieved by the proposed method against the PCC of the annotators (mental health experts), which has a mean value of **0.85** [5, 37] on NESS. We observe that the proposed method achieves a PCC close to that of human experts for the "Total Negative" and "EXP-Total" scores, in this dataset. In Fig. 3 we show the total score predictions for all videos, for both scales in the NESS dataset. As the NESS dataset is imbalanced, with fewer patients having severe symptoms, we observe a higher error for patients with higher ground truth labels. Moreover, since we perform a leave-one-patient-out cross-validation, there is a chance that no examples of high total scores are included in the training set of a given fold. This trend is consistent for both scales used to evaluate.

Table 6: Performance of proposed method against state-of-the-art methods on the PANSS-NEG symptom scale.

	N3: Poor Rapport			N6: Lack of Spontaneity			N1: Blunted Affect			Total Negative		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Tron *et al.* [44]	0.98	1.31	0.20	1.37	1.69	0.13	0.90	1.28	0.37	-	-	-
Tron *et al.* [45]	1.01	1.26	0.15	1.32	1.62	0.09	0.99	1.36	0.11	-	-	-
SchiNet [4]	**0.85**	1.20	0.27	1.25	1.51	0.25	0.84	1.18	0.42	3.30	4.17	0.29
Proposed	0.87	**1.16**	**0.78**	**0.74**	**0.95**	**0.47**	**0.64**	**0.87**	**0.56**	**2.80**	**3.78**	**0.71**

Table 7: Performance of proposed methodology against other state-of-the-art on the CAINS-EXP symptom scale.

	Facial Expression			Vocal Expression			Expressive Gestures			Quantity of Speech			EXP-Total Score		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Tron *et al.* [44]	0.80	1.03	0.37	0.87	1.23	0.23	0.85	1.19	0.36	1.09	1.43	0.27	-	-	-
Tron *et al.* [45]	0.75	1.07	0.36	0.86	1.22	0.26	0.91	1.22	0.38	1.02	1.36	0.25	-	-	-
SchiNet [4]	0.66	0.93	0.46	0.77	1.10	0.27	0.90	1.15	0.36	0.98	1.30	0.30	2.67	3.34	0.45
Proposed	**0.56**	**0.72**	**0.75**	**0.65**	**0.89**	**0.71**	**0.71**	**0.89**	**0.76**	**0.60**	**0.82**	**0.54**	**1.88**	**2.60**	**0.77**

Table 8: Performance of the proposed method against baseline and other uni-modal architectures (AMIGOS).

	Arousal		Valence	
	PCC	CCC	PCC	CCC
Proposed	**0.69**	**0.68**	**0.75**	**0.74**
Proposed $\mathcal{L}_{CCC}$ w/o K w/o $\mathcal{L}_{rel}$	0.59	0.49	0.64	0.54
Proposed $\mathcal{L}_{RMSE}$ w/o K w/o $\mathcal{L}_{rel}$	0.60	0.39	0.55	0.40
Mou *et al.* [35]	0.60	0.59	0.62	0.61

For experiments conducted on the OMG dataset [3], we compared the performance of the proposed method against other uni-modal multi-label works submitted to the "OMG-Emotion Behavior Challenge" – we show the results in Table 1, where we observe a clear improvement against previous works, for both Arousal and Valence in terms of CCC. We note that, to our knowledge, current state-of-the-art results for the OMG dataset are achieved by MIMAMO [12] with a CCC of 0.37 and 0.52 for Arousal and Valence respectively. However, as MIMAMO is a multi-modal approach (using RGB and inter-frame phase difference as input modalities) and is trained for a single target (i.e., Arousal or Valence) at a time, the results reported in [12] are not directly comparable to ours.

Finally, for the experiments conducted on the AMIGOS dataset [34], we compared the performance of the proposed methodology against previous state-of-the-art [35] for the face modality and we show the results in Table 8. The proposed methodology leads to a clear improvement against both baselines, trained with an RMSE regression loss ($\mathcal{L}_{RMSE}$) and a CCC loss ($\mathcal{L}_{CCC}$). We also outperform previous state-of-the-art by a large margin for both Arousal and Valence, even though we trained on a subset of the training data at each fold. It is worth noting that on the AMIGOS dataset the architecture that was pre-trained with a contrastive loss completely overfitted on the regression task and, thus, we choose to excluded it from the comparison. In Fig. 4, we see some visual examples of input clips, their context from the AMIGOS dataset [34] and the proposed methodology predictions against the ground truth.

6 CONCLUSION

In this work we presented our method for dealing with challenges that arise in the domain of affect and mental health in multi-label regression problems. Specifically, we built on [48] and proposed a two-stage attention architecture that uses features from the clips' neighbourhood to introduce context information in the feature extraction. The architecture is novel in the domain of affect and mental state analysis and leads to smaller training times in comparison to state of the art. Furthermore, we introduced a novel relational regression loss that aims at learning from the label relationships of the samples during training. The proposed loss uses the distance between label vectors to learn intra-batch latent representation similarities in a supervised manner. We showed that the improved latent representations obtained with the addition of the relational regression loss lead to improved regression output, without the use of large datasets. Finally, we demonstrated the effectiveness of the proposed method on three datasets for schizophrenia symptom severity estimation and for continuous affect estimation, and we showed that our method achieves results comparable to the state-of-the-art – specifically for symptom severity estimation of schizophrenia, our methodology outperforms the previous state-of-the-art on all scales and symptoms tested and achieves a Pearson's Correlation Coefficient similar to that of human experts.

ACKNOWLEDGMENTS

This work is supported by EPSRC DTP studentship (No. EP/R513106/1) and EU H2020 AI4Media (No. 951911).

REFERENCES

[1] Samuel Albanie and Andrea Vedaldi. 2016. Learning Grimaces by Watching TV. In *BMVC 2016*. http://arxiv.org/abs/1610.02255 arXiv: 1610.02255.

[2] American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders* (fifth edition ed.). American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

[3] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. 2018. The OMG-Emotion Behavior Dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*. 1–7. https://doi.org/10.1109/IJCNN.2018.8489099 ISSN: 2161-4407.

[4] Mina Bishay, Petar Palasek, Stefan Priebe, and Ioannis Patras. 2018. SchiNet: Automatic Estimation of Symptoms of Schizophrenia from Facial Behaviour Analysis. *IEEE Transactions on Affective Computing* (2018). https://doi.org/10.1109/TAFFC.2019.2907628

[5] Mina Adel Thabet Bishay. 2020. *Automatic Facial Expression Analysis in Diagnosis and Treatment of Schizophrenia.* Thesis. Queen Mary University of London. https://qmro.qmul.ac.uk/xmlui/handle/123456789/69449 Accepted: 2020-12-18T16:23:10Z.

[6] Adrian Bulat, Shiyang Cheng, Jing Yang, Andrew Garbett, Enrique Sanchez, and Georgios Tzimiropoulos. 2021. Pre-training strategies and datasets for facial representation learning. *arXiv:2103.16554 [cs]* (March 2021). http://arxiv.org/abs/2103.16554 arXiv: 2103.16554.

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 67–74.

[8] Wheidima Carneiro de Melo, Eric Granger, and Abdenour Hadid. 2020. A Deep Multiscale Spatiotemporal Network for Assessing Depression from Facial Dynamics. *IEEE Transactions on Affective Computing* (2020), 1–1. https://doi.org/10.1109/TAFFC.2020.3021755

[9] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A Short Note about Kinetics-600. *arXiv:1808.01340 [cs]* (Aug. 2018). http://arxiv.org/abs/1808.01340 arXiv: 1808.01340.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[11] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems* 33 (2020), 22243–22255.

[12] Didan Deng, Zhaokang Chen, Yuqian Zhou, and Bertram Shi. 2020. MIMAMO Net: Integrating Micro- and Macro-Motion for Video Emotion Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 03 (April 2020), 2621–2628. https://doi.org/10.1609/aaai.v34i03.5646

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[14] Paul Ekman and Wallace V. Friesen. 1978. *Facial action coding system: Investigator's guide*. Consulting Psychologists Press.

[15] Courtney Forbes, Jack J. Blanchard, Melanie Bennett, William P. Horan, Ann Kring, and Raquel Gur. 2010. Initial development and preliminary validation of a new negative symptom measure: The Clinical Assessment Interview for Negative Symptoms (CAINS). *Schizophrenia Research* 124, 1-3 (Dec. 2010), 36–42. https://doi.org/10.1016/j.schres.2010.08.039

[16] Niki Maria Foteinopoulou, Christos Tzelepis, and Ioannis Patras. 2021. Estimating continuous affect with uncertainty. Nara, Japan. https://doi.org/10.1109/ACII52823.2021.9597425

[17] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 315–323.

[18] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. 2013. Challenges in Representation Learning: A report on three machine learning contests. *arXiv:1307.0414 [cs, stat]* (July 2013). http://arxiv.org/abs/1307.0414 arXiv: 1307.0414.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[20] Tomu Hirata, Yusuke Mukuta, and Tatsuya Harada. 2021. Making Video Recognition Models Robust to Common Corruptions With Supervised Contrastive Learning. In *ACM Multimedia Asia (MMAsia '21)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3469877.3497692

[21] Yan-Jia Huang, Yi-Ting Lin, Chen-Chung Liu, Lue-En Lee, Shu-Hui Hung, Jun-Kai Lo, and Li-Chen Fu. 2022. Assessing Schizophrenia Patients through Linguistic and Acoustic Features using Deep Learning Techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2022), 1–1. https://doi.org/10.1109/TNSRE.2022.3163777 Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

[22] Jyoti Joshi, Roland Goecke, Sharifa Alghowinem, Abhinav Dhall, Michael Wagner, Julien Epps, Gordon Parker, and Michael Breakspear. 2013. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces* 7, 3 (2013), 217–228. https://doi.org/10.1007/s12193-013-0123-2

[23] Jyoti Joshi, Roland Goecke, Gordon Parker, and Michael Breakspear. 2013. Can Body Expressions Contribute to Automatic Depression Analysis? *Automatic Face and Gesture Recognition (FG), 2013 10Th IEEE International Conference and Workshops* (2013). https://doi.org/10.1109/FG.2013.6553796

[24] S. R. Kay, A. Fiszbein, and L. A. Opler. 1987. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bulletin* 13, 2 (Jan. 1987), 261–276. https://doi.org/10.1093/schbul/13.2.261

[25] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *arXiv:1705.06950 [cs]* (May 2017). http://arxiv.org/abs/1705.06950 arXiv: 1705.06950.

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 18661–18673. https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf

[27] Daeha Kim and Byung Cheol Song. 2021. Contrastive Adversarial Learning for Person Independent Facial Emotion Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 7 (May 2021), 5948–5956. https://ojs.aaai.org/index.php/AAAI/article/view/16743 Number: 7.

[28] Dimitrios Kollias and Stefanos Zafeiriou. 2019. A Multi-component CNN-RNN Approach for Dimensional Emotion Recognition in-the-wild. *arXiv:1805.01452 [cs, eess, stat]* (Dec. 2019). http://arxiv.org/abs/1805.01452 arXiv: 1805.01452.

[29] Fu Li, Chuang Gan, Xiao Liu, Yunlong Bian, Xiang Long, Yandong Li, Zhichao Li, Jie Zhou, and Shilei Wen. 2017. Temporal Modeling Approaches for Large-scale Youtube-8M Video Understanding. *arXiv:1707.04555 [cs]* (July 2017). http://arxiv.org/abs/1707.04555 arXiv: 1707.04555.

[30] Xuewei Li, Hongjun Wu, Mengzhu Li, and Hongzhe Liu. 2022. Multi-label video classification via coupling attentional multiple instance learning with label relation graph. *Pattern Recognition Letters* 156 (2022), 53–59. https://doi.org/10.1016/j.patrec.2022.01.003

[31] Cheng Lu, Wenming Zheng, Chaolong Li, Chuangao Tang, Suyuan Liu, Simeng Yan, and Yuan Zong. 2018. Multiple Spatio-Temporal Feature Learning for Video-Based Emotion Recognition in the Wild. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 646–652. https://doi.org/10.1145/3242969.3264992

[32] David Melhart, Antonios Liapis, and Georgios N. Yannakakis. 2021. Towards General Models of Player Experience: A Study Within Genres. In *2021 IEEE Conference on Games (CoG)*. 01–08. https://doi.org/10.1109/CoG52621.2021.9618902

[33] W. C. de Melo, E. Granger, and A. Hadid. 2019. Depression Detection Based on Deep Distribution Learning. In *2019 IEEE International Conference on Image Processing (ICIP)*. 4544–4548. https://doi.org/10.1109/ICIP.2019.8803467 ISSN: 2381-8549.

[34] Juan Abdon Miranda Correa, Mojtaba Khomami Abadi, Niculae Sebe, and Ioannis Patras. 2018. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing* (2018), 1–1. https://doi.org/10.1109/TAFFC.2018.2884461

[35] Wenxuan Mou, Hatice Gunes, and Ioannis Patras. 2019. Alone versus In-a-group: A Multi-modal Framework for Automatic Affect Recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 2 (June 2019), 1–23. https://doi.org/10.1145/3321509

[36] Songyou Peng, Le Zhang, Yutong Ban, Meng Fang, and Stefan Winkler. 2018. A deep network for arousal-valence emotion prediction with acoustic-visual cues. *arXiv preprint arXiv:1805.00638* (2018).

[37] Stefan Priebe, Mark Savill, Til Wykes, RP Bentall, Ulrich Reininghaus, Christoph Lauber, Stephen Bremner, Sandra Eldridge, and Frank Röhricht. 2016. Effectiveness of group body psychotherapy for negative symptoms of schizophrenia: multicentre randomised controlled trial. *The British Journal of Psychiatry* 209, 1 (2016), 54–61.

[38] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. 2021. Spatiotemporal Contrastive Video Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6964–6974.

[39] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178. https://doi.org/10.1037/h0077714

[40] Doyo Setiono, David Saputra, Kaleb Putra, Jurike V. Moniaga, and Andry Chowanda. 2021. Enhancing Player Experience in Game With Affective Computing. *Procedia Computer Science* 179 (2021), 781–788. https://doi.org/10.1016/j.procs.2021.01.066 5th International Conference on Computer Science and Computational Intelligence 2020.

[41] Erin Smith, Eric A. Storch, Helen Lavretsky, Jeffrey L. Cummings, and Harris A. Eyre. 2020. *Affective Computing for Brain Health Disorders*. Springer International Publishing, Cham, 1–14. https://doi.org/10.1007/978-3-319-75479-6_36-1

[42] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* 3, 1 (Jan. 2021), 42–50. https://doi.org/10.1038/s42256-020-00280-0 Number: 1 Publisher: Nature Publishing Group.

[43] Minh Tran, Ellen Bradley, Michelle Matvey, Joshua Woolley, and Mohammad Soleymani. 2021. Modeling Dynamics of Facial Behavior for Mental Health Assessment. (Aug. 2021). https://scirate.com/arxiv/2108.09934

[44] Talia Tron, Abraham Peled, Alexander Grinsphoon, and Daphna Weinshall. 2015. Automated facial expressions analysis in schizophrenia: A continuous dynamic approach. In *International Symposium on Pervasive Computing Paradigms for Mental Health*. Springer, 72–81.

[45] Talia Tron, Abraham Peled, Alexander Grinsphoon, and Daphna Weinshall. 2016. Facial expressions and flat affect in schizophrenia, automatic analysis from depth camera data. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 220–223.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]* (Dec. 2017). http://arxiv.org/abs/1706.03762 arXiv: 1706.03762.

[47] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. 2020. Multi-Label Classification with Label Graph Superimposing. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 07 (April 2020), 12265–12272. https://doi.org/10.1609/aaai.v34i07.6909

[48] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross Girshick. 2019. Long-Term Feature Banks for Detailed Video Understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019). https://openaccess.thecvf.com/content_CVPR_2019/html/Wu_Long-Term_Feature_Banks_for_Detailed_Video_Understanding_CVPR_2019_paper.html

[49] Elaheh Yadegaridehkordi, Nurul Fazmidar Binti Mohd Noor, Mohamad Nizam Bin Ayub, Hannyzzura Binti Affal, and Nornazlita Binti Hussin. 2019. Affective computing in education: A systematic review and future research. *Computers & Education* 142 (2019), 103649. https://doi.org/10.1016/j.compedu.2019.103649

[50] Zhengyuan Yang, Amanda Kay, Yuncheng Li, Wendi Cross, and Jiebo Luo. 2021. Pose-based Body Language Recognition for Emotion and Psychiatric Symptom Interpretation. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 294–301. https://doi.org/10.1109/ICPR48806.2021.9412591

[51] Li-Wei Zhang, Jingting Li, Su-Jing Wang, Xian-Hua Duan, Wen-Jing Yan, Hai-Yong Xie, and Shu-Cheng Huang. 2020. Spatio-temporal fusion for Macro- and Micro-expression Spotting in Long Video Sequences. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. 734–741. https://doi.org/10.1109/FG47880.2020.00037

[52] Jiaming Zhou, Kun-Yu Lin, Haoxin Li, and Wei-Shi Zheng. 2021. Graph-Based High-Order Relation Modeling for Long-Term Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8984–8993.