Supervised Contrastive Learning for Affect Modelling

Kosmas Pinitas
Institute of Digital Games, University of Malta
Msida, Malta
kosmas.pinitas@um.edu.mt

Konstantinos Makantasis
Institute of Digital Games, University of Malta
Msida, Malta
konstantinos.makantasis@um.edu.mt

Antonios Liapis
Institute of Digital Games, University of Malta
Msida, Malta
antonios.liapis@um.edu.mt

Georgios N. Yannakakis
Institute of Digital Games, University of Malta
Msida, Malta
georgios.yannakakis@um.edu.mt

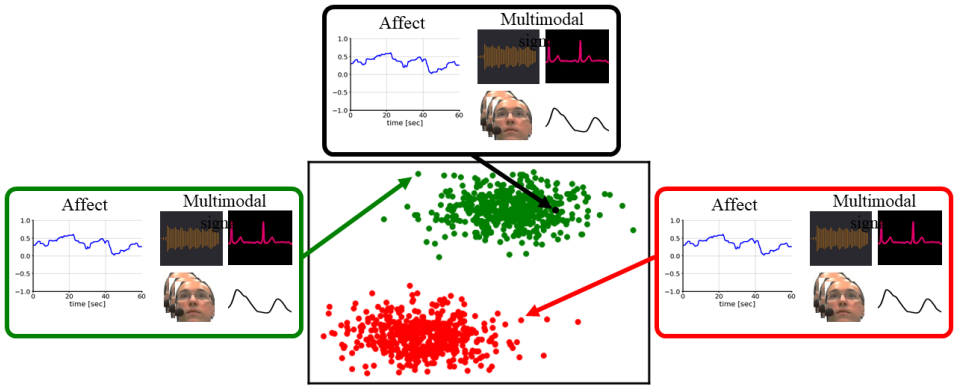


Figure 1: A high-level visualisation of the concept introduced. Supervised contrastive learning operates by infusing affect information within the representation, by pairing positive embeddings and dissociating negative embeddings. We assume affect is embedded in the multimodal latent space and defines what distinguishes (contrasts) data. Positive (green) and negative (red) multimodal data is labelled with respect to an anchor affect (black). Similar and dissimilar affect patterns define, respectively, positive and negative pairs of the anchor. The resulting representation is general with respect to affect patterns in a corpus.

ABSTRACT

Affect modeling is viewed, traditionally, as the process of mapping measurable affect manifestations from multiple modalities of user input to affect labels. That mapping is usually inferred through end-to-end (manifestation-to-affect) machine learning processes. What if, instead, one trains general, subject-invariant representations that consider affect information and then uses such representations to model affect? In this paper we assume that affect labels form an integral part, and not just the training signal, of an affect representation and we explore how the recent paradigm of contrastive learning can be employed to discover general high-level affect-infused representations for the purpose of modeling affect. We introduce three different supervised contrastive learning approaches for training representations that consider affect information. In this initial study we test the proposed methods for arousal prediction in the RECOLA dataset based on user information from multiple modalities. Results demonstrate the representation capacity of contrastive learning and its efficiency in boosting the accuracy of affect models. Beyond their evidenced higher performance compared to end-to-end arousal classification, the resulting representations are general-purpose and subject-agnostic, as training is guided though general affect information available in any multimodal corpus.

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556584

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence*; **Neural networks**; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

contrastive learning; affective computing; arousal; multimodal affect modeling

ACM Reference Format:
Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N. Yannakakis. 2022. Supervised Contrastive Learning for Affect Modelling. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536221.3556584

1 INTRODUCTION

Contrastive learning is a recent machine learning paradigm which has been widely and successfully employed for learning general representations of data [21, 38]. As a self-supervised learning method, it aims to project data into a space in which different views of the same input have similar representations [21]. Although contrastive learning is quite popular within the computer vision domain [9, 18], such methods have found applications in affective computing only recently for learning subject-invariant representations [43]. This paper builds on the hypothesis that affect information is an inherent property of the manifestations of affect and thus can be fused and learned in a contrastive learning manner, i.e. learning general affect-infused representations (see Fig. 1). In particular, we reframe the way multimodal affect modeling is viewed traditionally (end-to-end) and utilise affect annotations to build contrastive labels which are then used to train multimodal affect models.

To test our hypothesis that contrastive learning can learn general representations of affect and yield effective multimodal affect models, we build upon the Supervised Contrastive Learning (SCL) framework [19] for representation learning. In particular, we introduce and investigate three different approaches for building contrastive labels that rely on affect information (i.e. annotations). We evaluate the proposed methods against end-to-end classification using the extracted features of all modalities and corresponding arousal annotations of the RECOLA database [36]. The approach investigated in this initial study feeds features captured within a time window into a neural network model trained via SCL. A probe model [4] is then employed to assess the quality of the learned representations by predicting high and low arousal states (i.e. arousal classification). Results indicate that SCL yields arousal models that perform *significantly* better than end-to-end classification. This suggests that (a) contrastive learning is beneficial for downstream multimodal affect modelling tasks and (b) we can indeed fuse and blend affect information into the learnt representations through contrastive learning.

This paper is novel in several ways. First, to the best of our knowledge, this is the first time that the SCL framework has been employed in the context of affect modeling for learning affect-infused multimodal representations. Second, the generalization capacity of the representations learned via SCL is tested through three methods that utilise affect as a contrastive learning training signal. In particular, representations are trained on *affect classes*, *affect changes* and *affect trends*. Finally, the proposed approach is compared against end-to-end classification across dissimilar modalities of the RECOLA database for predicting arousal, thereby extending the relatively sparse volume of work on the intersection of contrastive learning and multimodal affective computing.

2 RELATED WORK

This section surveys related work on affect modelling using multiple and dissimilar user modalities, and moves on to review literature on the intersection of affect modelling and contrastive learning.

2.1 Multimodal Affect Modeling

As emotions can be manifested through various user modalities including physiology, speech (audio) and facial expression (video) [6, 34] it is no surprise that aforementioned modalities of user data have been used extensively for modeling affect. When it comes to affect modeling based solely on visual information, the dominant approach for years has been to leverage domain knowledge and consequently to manually author high-level hand-crafted visual features or to employ classic pattern recognition techniques [8, 51]. The advent of deep learning, however, allowed representation extraction to be automated and consequently revolutionized the field of pixel-based affect modeling. To the best of our knowledge, the study of Baveye et al. [3] is the first to test the effectiveness of deep learning in the context of emotion prediction in videos. The authors concluded that although the size of the dataset might prevent the effectiveness of deep learning, using representations based on convolutional neural networks (CNNs) as features is promising for affective movie content analysis. Since then, deep learning has been used to extract relevant representations in many applications of affect modeling. Indicatively, Ng et al. [33] used a deep CNN pre-trained on the generic ImageNet dataset and leveraged transfer learning techniques to perform emotion recognition on small datasets. Makantasis et al. [26] employed three dissimilar CNN architectures to predict player arousal from gameplay footage, showcasing that a mapping between gameplay video streams and the player's arousal exists.

The field of emotion recognition via audio advanced significantly by leveraging domain knowledge, hand-crafted audio features and classic pattern recognition methods [10, 39]. Nevertheless, deep learning algorithms quickly became a common practice in the field due to their high predictive capacity [1, 23]. Indicatively, Huang et al. [17] learned candidate features via contractive CNNs and fed the learned features to a semi-CNN in order to extract affect-salient features. The experimental results showcased that the affect-salient features outperformed well-established features in speech emotion recognition. Kwon [20] introduced a novel deep learning model for speech emotion recognition that utilizes a lightweight dilated CNN architecture and implemented the multi-learning trick approach. The model was evaluated on two benchmark datasets obtaining a high recognition accuracy.

Emotion can be effectively captured by a subject's physiological response to a set of stimuli since physiological reactions (e.g. changes in brain activity and in somatic and visceral systems) are measurable manifestations of affect [5]. Early work in the field of psychophysiology focused on the use of affect models that map between hand-authored physiological features and affect [16, 29]. Martinez et al. [30] introduced the first deep learning (CNN-based) approach for physiology-based affect modeling while Harper and Southern [15] presented an end-to-end deep learning model capable of classifying emotional valence from heartbeat time series along with a Bayesian framework for determining the uncertainty of the predictions. Giannakakis et al. [12] employed a multi-kernel 1D CNN to compute complex feature maps performing multi-level modeling of the unique heart rate variability signature for stress identification.

Fusing more than one user modalities into an affect model (multimodal affect modeling) has been an active research field [2, 42]; thus we will only focus on a few indicative studies. The work of Martinez et al. [31] is arguably one of the first investigating the fusion of user modalities via a deep learning perspective. Recently, Makantasis et

al. [27] evaluated the capacity of deep learned representations to predict affect by relying only on audiovisual information of videos. Tzirakis et al. [45] employed a CNN and a deep residual network of 50 layers to extract features from speech and video modalities showcasing that deep learning can outperform traditional approaches in terms of concordance correlation coefficient. Guo et al [14] compared four combinations of eye images, eye movement and electroencephalograms and two fusion methods, demonstrating that different modalities provide complementary information for recognizing five emotions. Zhang et al. [49] employed a Convolutional LSTM to extract spatio-temporal facial features and a 1D CNN to extract bio-sensing features. Finally, the works in [25, 28] explored the concept of privileged information for fusing information from multiple modalities into unimodal affect prediction models.

In contrast to all aforementioned studies, in this paper we derive high-level representations that are relevant for the affect modelling task at hand through contrastive learning. The predictive capacity of these representations is tested in the task of arousal classification across the different modalities of the RECOLA database. Unlike previous studies which used either pre-trained models or models trained specifically (end-to-end) for each task, we test an intermediate solution which account for the particularities of the dataset but produces representations of affect that can be re-used for different downstream tasks.

2.2 Contrastive Learning for Affect Modeling

While research at the intersection of affective computing and contrastive learning (CL) algorithms has been active over the last few years, the literature is still relatively sparse. CL framework which utilizes a spatiotemporal augmentation scheme for facial expression recognition in videos. The same authors exploited facial images captured simultaneously from different angles and developed a two-step training process to address the Multi-view Facial Expression Recognition problem [37]. Li et al. [22] investigated the impact of unsupervised representation learning on unlabeled datasets for speech emotion recognition and demonstrated that the contrastive predictive coding method can learn salient representations from unlabeled datasets that achieve state-of-the-art performance for the activation, valence and dominance primitives. Mai et al. [24] introduced a novel hybrid contrastive learning of tri-modal representation framework, using intra-/inter-modal and semi-contrastive learning to allow the model to explore cross-modal interactions and preserve inter-class relationships that reduce the modality gap. Finally, Yin et al. [48] proposed a two-step framework based on contrastive learning in order to address the cross-corpus emotion recognition problem. The authors demonstrate that utilizing contrastive learning to pre-train the encoder in a specific domain can produce representations that can be finetuned in a similar domain and achieve superior performance in the new domain.

In contrast to the aforementioned studies, this work contributes to the literature by introducing three different strategies for constructing the supervision signal used in SCL. All above-mentioned studies yield representations using solely information of the input space of the affect model. In this study, instead, we employ SCL to derive representations using affect-based contrastive labels. In other words, we fuse and blend affect information on the input space (i.e., subject measurements such as visual, audio and physiology data used in affect modelling) of learnt representations. Towards this direction, we exploit different statistical properties of continuous subjectively-defined affect annotations. Moreover, we investigate the impact of the different strategies on the expressiveness of the obtained representations and provide useful insights regarding their capacity to predict affective states. Results verify that employing contrastive learning representations boosts the performance of affect models for both uni- and multimodal affect measurements.

3 METHODOLOGY

This section describes the main elements of the algorithms examined in this paper. We start by presenting the representation learning components and move on with the methods employed for constructing the affect-based supervision signal for contrastive learning, i.e., contrastive labels, by selecting positive and negative samples based on continuous arousal annotation traces. The code for this paper is available on GitHub[1].

3.1 Representation Components

In this section we first present the main components used in representation learning, namely encoders and probes as depicted in Fig. 2. Then, we present a baseline architecture of an end-to-end classifier that we compare against our SCL methodology for assessing the effectiveness of the obtained affect infused representations.

3.1.1 Encoder. An encoder model E can be characterized by any neural network architecture that projects high dimensional data into a latent space of lower dimension, producing high-level representations of the input data. Hence, after training, E is an efficient coding function that reduces the dimensionality of the data while maintaining essential information about the input space. In this work, we hypothesize that affect information is a manifested and embedded property of the input space and consequently, it can be merged with the latent space via contrastive learning to yield more robust representations. We thus derive contrastive labels of affect and subsequently train the encoder using the L_{SCL} supervised contrastive loss function [19]:

$$L_{SCL} = \sum_{s \in S} \frac{-1}{|P_s|} \sum_{p \in P_s} \log \frac{\exp(r_s \cdot r_p / \tau)}{\sum_{a \in A_s} \exp(r_s \cdot r_a / \tau)} \quad (1)$$

where S is a set that includes all samples and P_s is the set that includes only the samples that are assigned to the same class as s (positive sample set). In addition, A_s is a set that contains any element of set S besides element s. With r_s, r_p and r_a we denote the hidden representations of the model for the samples s, p and a, respectively. Finally, τ stands for a non-negative temperature hyperparameter.

3.1.2 Probe. Probe architectures are mainly used to evaluate the quality of representations extracted from a pre-trained encoder E. In particular, given a known property (e.g. object categories in object recognition problems) of the input data, a probe can be trained to determine whether or not this property has been transferred to the latent space. Although there is no limitation in the number of

[1] https://github.com/kpinitas/Supervised-Contrastive-Learning-for-Affect-Modeling

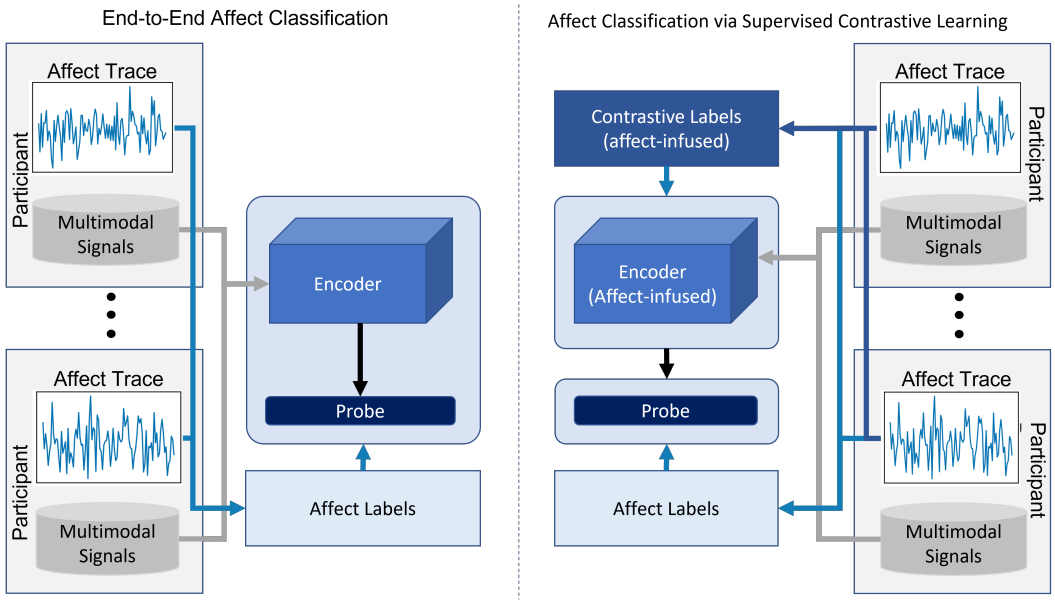


Figure 2: Illustration of the training methods employed: the end-to-end classification baseline (left) and the SCL method (right) which first derives affect-infused labels for contrastive learning, and then trains the probe model based on the affect representations of the pre-trained encoder. In both learning paradigms, affect labels are derived from participants' annotations.

hidden layers, a probe architecture usually consists of a single layer with softmax activation function, i.e. a *linear probe* [4].

3.1.3 Baseline. The baseline architecture used in this work realizes end-to-end classification (without the representation learning carried out via contrastive labels discussed in Section 3.2) and is labeled E_b. The model for E_b consists of a randomly initialized encoder E followed by a randomly initialized probe architecture. Both the encoder and the probe are trained (simultaneously) to map high-level handcrafted multimodal features to affect labels.

3.2 Affect-Infused Contrastive Labels

The core contribution of this paper is the use of supervised contrastive learning algorithms for yielding general affect-infused representations that would be beneficial to any downstream affect modelling task. Specifically, we use SCL to pre-train the encoder architecture (see Section 3.1.1) using affect-based contrastive labels to fuse affect information into representation learning. Then, we freeze the encoder's weights and train only the linear probe architecture (see Section 3.1.2) according to the downstream task. This training process is depicted in Fig. 2.

Arguably one of the most crucial steps in contrastive learning is the choice of positive and negative samples, which highly affects the quality of the CL supervision signal. The selection process is relatively straightforward when the annotations can be defined objectively (e.g. class labels) since a positive pair usually consists of two samples of the same class and a negative pair consists of samples that belong to different classes [19]. However, when it comes to subjective annotations such as arousal signals, there is no clear pair selection strategy due to their subjective nature and data noise caused by the inherent bias of human annotators [47].

This study explores three dissimilar affect measurements that can be calculated from any continuous annotation trace: one *absolute* measure regarding the subject's current emotional state per se and two *relative* measures regarding how the continuous emotional state fluctuates during a predefined time window [7, 47]. We calculate the *affect state* (g_a) as the mean of the affect trace captured within a time window (Eq. 2) [27, 32], the *affect change* score (c_a) as the average of the absolute differences between consecutive annotation values (Eq. 3) and the *affect trend* score (t_a) as the average of the differences between consecutive annotation values (Eq. 4):

$$g_a = \frac{1}{w}\sum_{i=0}^{w} v_i \tag{2}$$

$$c_a = \frac{1}{w}\sum_{i=1}^{w} |v_i - v_{i-1}| \tag{3}$$

$$t_a = \frac{1}{w}\sum_{i=1}^{w} (v_i - v_{i-1}) \tag{4}$$

where w is the window size considered and v_i is the i-th annotation value of the time window.

Based on the aforementioned measures, we explore three different positive/negative sample selection strategies that we detail below. It should be noted that, for all the proposed contrastive labelling strategies, we train the SCL models using the same loss function: i.e. the supervised contrastive loss function L_{SCL} of Eq. (1).

3.2.1 Contrasting Affect: High vs. Low . Intuitively, contrastive labels can be constructed by matching windows that have similar affect states as positive pairs and those with dissimilar affect states as negative pairs. To define affect state similarity, we binarize affect

states g_a's as "high" and "low", and consider windows with same (different) states as similar (dissimilar). The binarization criterion is based on the median ground truth value of the entire set of affect annotation traces ($\tilde{g}_a$) and on a threshold ϵ. Specifically, a time window i is labeled as "high" when $g_{a_i} > \tilde{g}_a + \epsilon$ and as "low" when $g_{a_i} < \tilde{g}_a - \epsilon$. It should be noted that the threshold ϵ—as e.g. employed in [26]—is used to eliminate windows with ambiguous affect annotation values close to the median which may deteriorate the stability of the SCL models and thus the effectiveness of the learned representations. The resulting preprocessed dataset that does not include the ambiguous affect values forms the basis for all three labeling strategies.

3.2.2 Contrasting Affect: Change vs. Unchanged . While the high-low pairing strategy of Section 3.2.1 uses an absolute measure of affect, a similar binarization procedure can be performed based on *affect change* (see Eq. 3) which is a relative measure. We can label a time window i as "change" when $c_{a_i} > \tilde{c}_a$ and as "unchanged" (i.e. no change) when $c_{a_i} \leq \tilde{c}_a$. Once again, we set $\tilde{c}_a$ to be the median affect change value of the entire set of affect change traces. By selecting the median $\tilde{c}_a$ to binarize the data, we end up with a balanced dataset. As in Section 3.2.1, we use these labels to pair windows i and j as positive when they both feature affect change or both feature no change, and as negative when one window features affect change and the other one does not.

3.2.3 Contrasting Affect: Uptrend vs. Downtrend. Inspired by [47] and similarly to the pairing strategy of Section 3.2.2, this contrastive labelling strategy uses a relative measure of affect, that of *affect trend*. Hence the i-th time window is assigned to the "uptrend" class when $t_{a_i} > \tilde{t}_a$ otherwise it is assigned to the "downtrend" class. We set $\tilde{t}_a$ to be the median affect trend value of the entire set of affect trend traces. Once again, we use the labels to define positive and negative samples based on the class that they belong to: a class match and a class mismatch define positive and negative pairs, respectively.

The main difference between the first and the other two contrastive labelling strategies is that the former is direct as the "high" and "low" values are derived from the actual magnitude of the affect annotation trace (see Eq. 2). The other two strategies, instead, are indirect since both the "change" and the "trend" are higher order traces that express the average absolute rate of change (Eq. 3) and bent (Eq. 4) of the annotation trace, respectively. The binarization criterion for all three strategies, however, considers all the annotation traces of the affect corpus.

4 THE RECOLA DATABASE

The methodology proposed in this paper is tested on a challenging dataset of online dyadic interactions between 23 participants that includes recordings and emotion annotations which are part of the RECOLA Database [36]. RECOLA is a multimodal dataset that consists of audio, visual, and physiological recordings such as electrodermal activity (EDA), and electrocardiography (ECG). Aiming to strike a balance between the quantity and the quality of the features, RECOLA provides 40 handcrafted features for the video information, 116 for physiological signals and 130 for audio information. The video features correspond to action units, head pose estimation, and texture and optical flow features. The audio features are voicing related descriptors similar to those used in the COMPARE challenge [40, 41]. The physiological features extracted from EDA are related to the variability and zero-crossing of the signal as well as the spectral entropy and spectral coefficients. The ECG features instead include the spectral entropy and mean frequency of the signal.

The collected data is annotated by six assistants (i.e. annotators) in terms of arousal and valence. The annotations are continuous, bounded in the range of $[-1, 1]$, and provided at 25Hz. Apart from the features extracted from ECG, EDA signals, and raw footage information, the creators of the dataset also provide the videos from which the audiovisual features have been extracted. In this initial study, however, we do not make use of video frames (pixel information). It should be noted that amongst the 34 participants who gave their consent to share their data outside of the consortium, only 23 video recordings are publicly available. RECOLA has been used for audiovisual emotion recognition challenges in which the remaining 11 participants serve as an evaluation set, and thus their data is not publicly available.

4.1 Data Preprocessing

The RECOLA database provides both arousal and valence annotations. This dataset has been a benchmark for the AVEC challenge for several years. Consequently, the Hall of Fame[2] results can provide valuable insights into the quality of the features in predicting affect. In particular, the above results show that arousal is better captured in the provided audio, visual, and physiology features obtained by concatenating ECG and EDA features [35] [50] [46]. For this reason, in this initial study, we focus on the comparison among the introduced SCL variants (regarding arousal) in deriving robust contrastive learning representations for arousal modeling. In particular, we aim to derive informative representations that capture the temporal information encoded in the provided features and use them to predict arousal states.

Towards this direction, we split each participant's session (features) into overlapping time windows using a sliding step of 400ms and window lengths of 1, 2, 3 and 4 seconds. The sliding step and window length are hyperparameters that affect the size of the dataset and the information contained in each window, respectively. It should be noted that the features and the arousal annotations are already synchronized, and we do not need to account for the reaction time between stimulus and emotional response. After splitting each session into time windows, each window consists of a sequence of feature vectors. To reduce the computational load, we compute the average value for each feature inside the time window representing this way each time window by a single feature vector. Moreover, in this way the dimension of the feature vector is not dependant on the windows' length. Table 1 presents the number of samples per window length.

When it comes to affect annotation, we use the median annotation values per time window in order to mitigate inter-annotator disagreement [13]. The arousal state score g_a is computed based on Eq. (2) using the median arousal trace while c_a and t_a are computed according to Eq. (3) and (4) where the consecutive value differences

[2]Hall of Fame models: https://diuf.unifr.ch/main/diva/recola/news.html

Table 1: Number of samples obtained by applying different time window lengths

window length	**1 sec**	**2 sec**	**3 sec**	**4 sec**
sample size	10235	10124	9963	9811

correspond to arousal value differences. The g_a score is bound within $[-1, 1]$ as it captures the original scale of arousal annotation, and measures the general arousal level within the boundaries of the time window in question. The c_a score is zero when the annotation value remains constant throughout the time window (i.e. no reported change in arousal) and is high when the annotation changes drastically (regardless of whether it increases, decreases, or both). Finally, t_a is high when the annotation increases throughout the time window duration and low when the arousal score decreases. Note that unlike g_a, both c_a and t_a are unbounded, although in practice their values tend to be small.

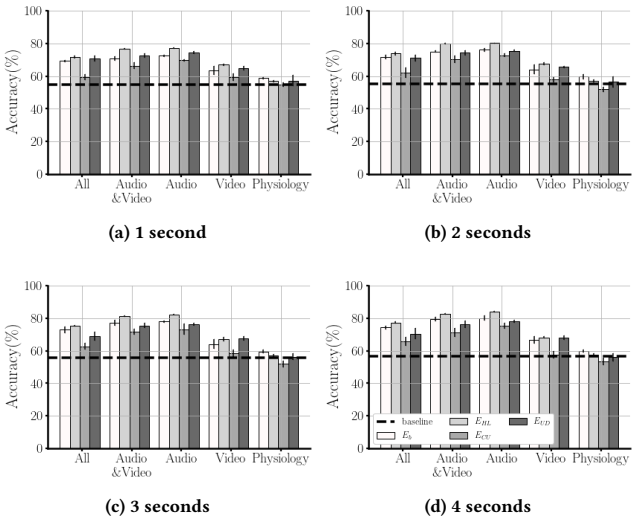
5 RESULTS

This section first outlines the experimental protocol we use for the evaluation of the methods introduced in this paper and then presents the key experimental results obtained.

5.1 Experimental Protocol

In this initial study, we test the proposed methods on the downstream task of arousal classification: the model has to learn to classify features within a time window as low or high arousal state. The encoder used in this work (E) is a simple ANN model with one sigmoid activated dense layer of 30 trainable neurons. The output of the dense layer, given a time window, corresponds to the high-level representation describing this specific window. The probe architecture is a dense layer of two neurons activated using the softmax function. Finally, our baseline model E_b is an encoder E followed by a probe. Each model in this paper is trained using the Adam optimizer with learning rate of 0.001 and batch size of 256. We set the temperature parameter τ in Eq. (1) to $\tau = 0.1$.

To generate the corresponding class labels ("high" vs. "low" arousal) we follow the processes described in Section 3.2.1. Moreover, the arousal change and arousal trend contrastive labels are generated only for the training samples of the downstream task to guarantee that the same training and test data are used for all models promoting a fair comparison among the presented algorithms. Furthermore, to evaluate the performance of each method, we use a five-fold cross-validation strategy for splitting the data into training and test sets ensuring that data in each set belong to different participants and thus the training and test sets are non-overlapping. The models are trained based on a convergence protocol that (early) stops the training process after 10 epochs without a training loss improvement and returns the model. The selected class splitting criterion (median of the arousal trace) and threshold $\epsilon = 0.1$ ensure that both the training and test sets are balanced and consequently, the performance of the models is evaluated in terms of accuracy score. We present experiments for time window lengths of 1, 2, 3 and 4 seconds.



(a) 1 second **(b) 2 seconds**

(c) 3 seconds **(d) 4 seconds**

Figure 3: Average 5-fold validation accuracy scores (%) for high-low arousal classification as a downstream task. Values are averaged across 10 independent runs; 95% confidence intervals are displayed as error bars.

5.2 Contrastive Learning For Arousal Modeling

We wish to investigate how representations learned via contrastive learning perform for a downstream task classifying between high and low arousal states. We thus train three encoders via SCL as per our three contrastive labeling strategies: high-low arousal state (E_{HL}), arousal change-unchanged (E_{CU}), and uptrend-downtrend (E_{UD}). We train a probe model for each of the encoders as mentioned in Section 3. The baseline model, E_b, performs end-to-end arousal classification. An additional baseline always chooses the most frequent class in the training set (dotted line in Fig. 3).

We explore five modality configurations of the RECOLA dataset: the single modalities of audio and video, the bimodal physiological signal containing ECG and EDA, the bimodal audio and video, and finally all modalities combined. Experiments across the four different modeling approaches, four time window lengths and five modality configurations are illustrated in Fig. 3. Results are based on 5-fold validation accuracy values, averaged across 10 independent runs. Statistical significance is established via a two-tailed Student's t-test with a significance threshold $p < 0.05$.

Figure 3 indicates that E_{HL} results in the best performing model. This is not surprising as the supervision signal in contrastive learning is the same as the downstream task. Specifically, the E_{HL} model outperforms the baseline end-to-end classifier (E_b) yielding significantly higher accuracy scores across all window lengths and RECOLA modalities except physiology. Results also showcase that the E_{UD} model performs on par and in some cases outperforms E_b, even though the contrastive learning supervision signal is different than the labels of the downstream task. Specifically, E_{UD} outperforms E_b significantly in 7 of 20 setups; for video input it outperforms E_b on all time windows. However, E_{UD} usually marks significantly lower accuracy scores than the E_{HL} model (in 16 of

20 setups). The E_{CU} model seems to perform poorly, reaching significantly lower accuracies than the E_b model in all cases. This indicates that some distinctions (change-unchanged) for representation learning are too distant to the downstream task to be useful.

Comparing across RECOLA modalities, the models achieve the highest accuracy when arousal modeling relies on audio features across all time window lengths. High accuracy scores are also obtained by the models when the audio features are fused in the input space with video features (Audio & Video) and with video and physiology (All) features. Although video features can yield robust arousal predictors, resulting models are inferior to the audio-based models. Finally, regardless of training method, all models underperform (at the same level as the baseline) when physiological signals are considered. It appears that arousal is not well manifested (or captured by) physiology in the RECOLA dataset.

Analysing some indicative key performance values obtained, it seems that E_{HL} achieves the highest average accuracy (i.e. 83.9%) and highest best-fold accuracy (i.e. 87.6%) across ten independent runs when using audio features as input, for 4 second time windows. This corresponds to 4.7% relative increase in accuracy compared to E_b. A similar pattern can be observed in the case of video-based arousal modeling where the highest average accuracy of 68.2% (and best-fold accuracy 73.8%) is achieved by E_{HL} (a 2.3% relative increase over E_b). It is worth noting that the performance of the models decreases for shorter time windows, but E_{HL} retains the best accuracies in all cases except physiological data; indicatively with audio features E_{HL} reaches an average accuracy of 77.2% (a 6.5% relative increase over E_b) on 1-second time windows. In stark contrast, when the models are trained solely on physiological features their performance is independent of the time window length as all models reach baseline performance. Physiological information seems irrelevant for the contrastive learning process as all SCL models perform worse than E_b.

6 DISCUSSION

This work investigated the potential of contrastive learning for handling affect modeling tasks when the annotations are continuous and subjectively defined. We aimed to assess the quality of feature-based representations of affect learned via supervised contrastive learning by comparing the performance of the learned representations with the representations learned via traditional arousal classification (high-low arousal) across the different modalities of the RECOLA Database. The results indicate that the SCL encoders yield more reliable models of affect when the affect modeling task is treated as a classification task.

A worthwhile discussion is our choice of not applying any data augmentation [44]. Although data augmentation is arguably a standard practice and consequently an integral part of the contrastive learning pipeline, we decided to omit this step since our input space in this initial study consists of hand-crafted features. Moreover, while data augmentation is prevalent in unsupervised contrastive learning, in our case we use affect labels for deriving positive pairs in a supervised fashion. It is worth exploring, however, whether additional data augmentation based on simple feature manipulation can augment the dataset and improve the models' performance.

In terms of future research, there are several directions that we can follow. An obvious next step towards generality is to test the efficiency of our model in predicting other affect dimensions such as valence or other core affective states such as happiness and fear. In addition, we plan to investigate the performance of our method in producing representations of affect when the affect modeling objective is treated as a regression or a preference learning task [11, 47]. Fine-tuning a large pre-trained model is a standard practice in contrastive learning when the model operates on the pixel-level of the image. Hence, another obvious next step for this work is to consider pixel-based information of image modalities and employ the proposed approach as a fine-tuning method for models that have already been trained on vast datasets (e.g. the ImageNet 1k dataset or more relevant face datasets). We did not use pixel-based representations and such pre-trained visual models in this initial study as we wanted to better investigate how the method performs with a simpler network and particularly compare it against an end-to-end training baseline. Finally, we note that the proposed methodology is general and thus applicable to any affective computing and classification task, as long as affect annotations exist and can be processed as labels.

7 CONCLUSIONS

This paper introduced a representation (contrastive) learning method that views affect as a training signal and integral part of the learned representation. We presented three approaches for handling subjectively defined continuous annotations, realising supervised contrastive learning in the domain of affective computing. Our experiments showcased that it is possible to learn highly-performing general affect-infused representations from arousal annotations of the RECOLA dataset. Comparing our method against end-to-end classification—which is one of the standard learning paradigms for modeling affect—we observe that some of the proposed SCL methods lead to significant improvements in performance which, in turn, showcases that our approach yields more accurate and reliable models of affect. While this first demonstration of supervised contrastive learning for affect-based representation learning tasks already shows a high potential for affect modeling, additional experiments considering more tasks, datasets, and learning paradigms are needed to assess the capacity and efficacy of the proposed approach.

ACKNOWLEDGMENTS

Kosmas Pinitas, Antonios Liapis and Georgios N. Yannakakis were supported by the European Union's H2020 research and innovation programme (Grant Agreement No. 951911). Konstantinos Makantasis was supported by the European Union's H2020 research and innovation programme (Grant Agreement No. 101003397).

REFERENCES

[1] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. 2021. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors* 21, 4 (2021).

[2] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends* 2, 2 (2021), 52–58.

[3] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, and Liming Chen. 2015. Deep learning vs. kernel methods: Performance for emotion prediction in

videos. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. IEEE, Xi'an, 77–83.
[4] Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *arXiv e-prints* (2021), arXiv–2102.
[5] Margaret M Bradley and Peter J Lang. 2000. Measuring emotion: Behavior, feeling, and physiology. In *Cognitive neuroscience of emotion*. Oxford University Press, 242–276.
[6] Rafael A Calvo, Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas. 2015. *The Oxford handbook of affective computing*. Oxford Library of Psychology.
[7] Elizabeth Camilleri, Georgios N. Yannakakis, and Antonios Liapis. 2017. Towards General Models of Player Affect. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. IEEE.
[8] Mohamed Dahmane and Jean Meunier. 2011. Emotion recognition using dynamic grid-based HoG features. In *Proceedings of the International Conference on Automatic Face & Gesture Recognition*. IEEE, 884–888.
[9] Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. 2021. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the International Conference on Computer Vision*. IEEE, 1502–1512.
[10] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition* 44, 3 (2011), 572–587.
[11] Johannes Fürnkranz and Eyke Hüllermeier. 2011. Preference learning. In *Encyclopedia of Machine Learning*. Springer, 789–795.
[12] Giorgos Giannakakis, Eleftherios Trivizakis, Manolis Tsiknakis, and Kostas Marias. 2019. A novel multi-kernel 1D convolutional neural network for stress recognition from ECG. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*. IEEE.
[13] Oliver Grewe, Frederik Nagel, Reinhard Kopiez, and Eckart Altenmüller. 2007. Emotions over time: synchronicity and development of subjective, physiological, and facial affective reactions to music. *Emotion* 7, 4 (2007), 774.
[14] Jiang-Jian Guo, Rong Zhou, Li-Ming Zhao, and Bao-Liang Lu. 2019. Multimodal emotion recognition from eye image, eye movement and EEG using deep neural networks. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 3071–3074.
[15] Ross Harper and Joshua Southern. 2020. A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. *IEEE Transactions on Affective Computing* 13, 2 (2020), 985–991.
[16] Christoffer Holmgård, Georgios N Yannakakis, Hector P Martinez, and Karen-Inge Karstoft. 2015. To rank or to classify? Annotating stress for reliable PTSD profiling. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*.
[17] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. 2014. Speech emotion recognition using CNN. In *Proceedings of the International Conference on Multimedia*. Association for Computing Machinery, 801–894.
[18] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2020), 2.
[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems* 33 (2020), 18661–18673.
[20] Soonil Kwon et al. 2021. MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Systems with Applications* 167 (2021), 114–177.
[21] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access* 8 (2020), 193907–193934.
[22] Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke, Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, and Chao Wang. 2021. Contrastive unsupervised learning for speech emotion recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6329–6333.
[23] Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. 2021. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics* 10, 10 (2021).
[24] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2021. Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis. *arXiv preprint arXiv:2109.01797* (2021).
[25] Konstantinos Makantasis. 2021. Affranknet+: ranking affect using privileged information. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*. IEEE.
[26] Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2019. From pixels to affect: A study on games and player experience. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. IEEE.
[27] Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2021. The pixels and sounds of emotion: General-purpose representations of arousal in games. *IEEE Transactions on Affective Computing* (2021).
[28] Konstantinos Makantasis, David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2021. Privileged Information for Modeling Affect In The Wild. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. IEEE.
[29] Regan L Mandryk and M Stella Atkins. 2007. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies* 65, 4 (2007), 329–347.
[30] Hector P Martinez, Yoshua Bengio, and Georgios N Yannakakis. 2013. Learning deep physiological models of affect. *IEEE Computational Intelligence magazine* 8, 2 (2013), 20–33.
[31] Héctor P Martínez and Georgios N Yannakakis. 2014. Deep multimodal fusion: Combining discrete events and continuous signals. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 34–41.
[32] David Melhart, Antonios Liapis, and Georgios N. Yannakakis. 2022. The Arousal video Game AnnotatIoN (AGAIN) Dataset. *IEEE Transactions on Affective Computing* (2022).
[33] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the International Conference on Multimodal Interaction*. Association for Computing Machinery, 443–449.
[34] Rosalind W Picard. 2000. *Affective computing*. MIT press.
[35] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters* 66 (2015), 22–30.
[36] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proceedings of the 10th International Conference and Workshops on Automatic Face and Gesture*. IEEE.
[37] Shuvendu Roy and Ali Etemad. 2021. Self-supervised contrastive learning of multi-view facial expressions. In *Proceedings of the International Conference. on Multimodal Interaction*. Association for Computing Machinery, 253–257.
[38] Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 3875–3879.
[39] Björn Schuller, Gerhard Rigoll, and Manfred Lang. 2003. Hidden Markov model-based speech emotion recognition. In *Proceedings of the International Conference on Multimedia and Expo*. IEEE.
[40] Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. 2014. The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load, multitasking. In *Proceedings of the Annual Conference of the International Speech Communication Association*.
[41] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings of the Annual Conference of the International Speech Communication Association*.
[42] Nicu Sebe, Ira Cohen, and Thomas S Huang. 2005. Multimodal emotion recognition. In *Handbook of pattern recognition and computer vision*. World Scientific, 387–409.
[43] Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. 2022. Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing* (2022).
[44] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Big Data* 6, 1 (2019).
[45] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1301–1309.
[46] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*. Association for Computing Machinery, 3–10.
[47] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. 2018. The ordinal nature of emotions: An emerging approach. *IEEE Trans. on Affective Computing* 12, 1 (2018), 16–35.
[48] Yufeng Yin, Liupei Lu, Yao Xiao, Zhi Xu, Kaijie Cai, Haonan Jiang, Jonathan Gratch, and Mohammad Soleymani. 2021. Contrastive Learning for Domain Transfer in Cross-Corpus Emotion Recognition. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. IEEE.
[49] Yuhao Zhang, Md Zakir Hossain, and Shafin Rahman. 2021. DeepVANet: a deep end-to-end network for multi-modal emotion recognition. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. Association for Computing Machinery, 227–237.

[50] Zixing Zhang, Jing Han, Eduardo Coutinho, and Björn Schuller. 2018. Dynamic difficulty awareness training for continuous emotion prediction. *IEEE Transactions on Multimedia* 21, 5 (2018), 1289–1301.

[51] Wenming Zheng, Hao Tang, Zhouchen Lin, and Thomas S Huang. 2010. Emotion recognition from arbitrary view facial images. In *Proc. of the European Conference on Computer Vision*. Springer, 490–503.