Weakly-supervised Learning for Fine-grained Emotion Recognition using Physiological Signals

Zhang, Tianyi; El Ali, Abdallah; Wang, Chen; Hanjalic, Alan; Cesar, Pablo

DOI
10.1109/TAFFC.2022.3158234

Publication date
2023

Document Version
Final published version

Published in
IEEE Transactions on Affective Computing

Citation (APA)
Zhang, T., El Ali, A., Wang, C., Hanjalic, A., & Cesar, P. (2023). Weakly-supervised Learning for Fine-grained Emotion Recognition using Physiological Signals. *IEEE Transactions on Affective Computing*, *14*(3), 2304-2322. https://doi.org/10.1109/TAFFC.2022.3158234

Weakly-Supervised Learning for Fine-Grained Emotion Recognition Using Physiological Signals

Tianyi Zhang, *Student Member, IEEE*, Abdallah El Ali, *Member, IEEE*, Chen Wang, Alan Hanjalic, *Fellow, IEEE*, and Pablo Cesar, *Senior Member, IEEE*

Abstract—Instead of predicting just one emotion for one activity (e.g., video watching), fine-grained emotion recognition enables more temporally precise recognition. Previous works on fine-grained emotion recognition require segment-by-segment, fine-grained emotion labels to train the recognition algorithm. However, experiments to collect these labels are costly and time-consuming compared with only collecting one emotion label after the user watched that stimulus (i.e., the post-stimuli emotion labels). To recognize emotions at a finer granularity level when trained with only post-stimuli labels, we propose an emotion recognition algorithm based on Deep Multiple Instance Learning (*EDMIL*) using physiological signals. *EDMIL* recognizes fine-grained valence and arousal (V-A) labels by identifying which instances represent the post-stimuli V-A annotated by users after watching the videos. Instead of fully-supervised training, the instances are weakly-supervised by the post-stimuli labels in the training stage. The V-A of instances are estimated by the instance gains, which indicate the probability of instances to predict the post-stimuli labels. We tested *EDMIL* on three different datasets, *CASE*, *MERCA* and *CEAP-360VR*, collected in three different environments: desktop, mobile and HMD-based Virtual Reality, respectively. Recognition results validated with the fine-grained V-A self-reports show that for subject-independent 3-class classification (high/neutral/low), *EDMIL* obtains promising recognition accuracies: 75.63% and 79.73% for V-A on *CASE*, 70.51% and 67.62% for V-A on *MERCA* and 65.04% and 67.05% for V-A on *CEAP-360VR*. Our ablation study shows that all components of *EDMIL* contribute to both the classification and regression tasks. Our experiments also show that (1) compared with fully-supervised learning, weakly-supervised learning can reduce the problem of overfitting caused by the temporal mismatch between fine-grained annotations and physiological signals, (2) instance segment lengths between 1-2 s result in the highest recognition accuracies and (3) *EDMIL* performs best if post-stimuli annotations consist of less than 30% or more than 60% of the entire video watching.

Index Terms—Deep multiple instance learning, emotion recognition, physiological signals, temporal ambiguity, weakly-supervised learning

1 Introduction

Recent years have witnessed a growing number of emotion recognition algorithms [1], [2], [3], [4] that particularly focus on modeling the temporal dynamics of emotion states. Recognizing users' emotion while they consume different types of media content (e.g., videos, music, movies) can help content providers better understand users' emotions towards their products and adjust the content accordingly [5]. For example, by identifying the moments that trigger negative emotions (e.g., confusion), the movie directors can improve the story arch of the narrative for their films and delete or adjust the scenes which make the audience distracted or confused. This requires techniques which can recognize emotions at a finer level of granularity, normally 0.5s to 4s according to prior emotion duration measures [3], [6], [7]. Compared with recognizing only one emotion for one video, fine-grained emotion recognition is temporally more precise as it can capture the time-varying nature of human emotions [8], [9], [10]: the emotions of users normally change continuously while watching videos.

Previous works [4], [10], [11] employ sequential machine learning algorithms such as Long Short Term Memory (LSTM) networks [12] to model the relationship between input signals and emotion states. However, sequential learning algorithms require fine-grained emotion labels for training. Here, the emotion labels and the input signals are required to have the same dimensions to train the recurrent structure of sequential learning algorithms [13]. To collect such fine-grained emotion labels (e.g., valence and arousal), there are typically three kinds of methods: (1) interrupt users at a fixed frequency for annotation [14], (2) ask users to annotate their emotions in real-time while watching videos [9], [15] or (3) let external observers annotate users' emotions segment-by-segment (e.g., using videos of users' facial expressions [16]) after watching videos [16], [17], [18]. However each of those methods has limitations. Requesting users to continuously annotate their emotions may not be feasible for longer durations (e.g., two hour film) as it may result in participant fatigue. Continuously interrupting people to self-report their emotional states can disrupt users' tasks [19]. For external

- *Tianyi Zhang and Pablo Cesar are with Distributed and Interactive Systems, Centrum Wiskunde & Informatica (CWI), 1098XG Amsterdam, The Netherlands, and also with Multimedia Computing Group, Delft University of Technology, 2600AA Delft, The Netherlands. E-mail: {tianyi, p.s.cesar}@cwi.nl.*
- *Abdallah El Ali is with Distributed and Interactive Systems, Centrum Wiskunde & Informatica (CWI), 1098XG Amsterdam, The Netherlands. E-mail: abdallah.elali@gmail.com.*
- *Chen Wang is with the State Key Laboratory of Media Convergence Production Technology and Systems, Xinhua News Agency & Future Media and Convergence Institute, Beijing 100031, China. E-mail: wangchen@news.cn.*
- *Alan Hanjalic is with Multimedia Computing Group, Delft University of Technology, 2600AA Delft, The Netherlands. E-mail: a.hanjalic@tudelft.nl.*

Manuscript received 14 Mar. 2021; revised 2 Jan. 2022; accepted 1 Mar. 2022. Date of publication 10 Mar. 2022; date of current version 13 Sept. 2023. (Corresponding author: Tianyi Zhang.) Recommended for acceptance by E. P. Scilingo. Digital Object Identifier no. 10.1109/TAFFC.2022.3158234

observers, some emotional states are difficult or misleading for them to annotate. For example, according to the experiments of Song *et al.* [20] and Abdic *et al.* [21], negative valence is often misidentified by external annotators as positive when users smile because of sarcasm and frustration. Even if collecting fine-grained emotion labels is possible, the experiments to collect them are time-consuming and costly [15]. Researchers have to spend extra time and money collecting fine-grained emotion labels because it is an additional task other than the data collection experiment (e.g., asking users to (re-)watch videos).

Given these issues, it is no surprise that a large number of datasets [22], [23], [24] that only contain one emotion label annotated by the user after watching one video stimulus (i.e., the post-stimuli labels). Instead of multiple labels for every fine-grained time segment (instance), there is only one post-stimuli label for one activity (i.e., one user watches one video). However, according to the *peak-end theory* [25], the post-stimuli labels represent only the most salient (peak) or recent (end) emotion during the video watching rather than the naturally dynamic and subtle emotional changes that may occur within it. According to Romeo *et al.* [19], this is defined as the problem of *temporal ambiguity*. When training machine learning algorithms to recognize fine-grained emotions using post-stimuli labels, the information on which fine-grained instances represent the emotion users labeled post-stimuli is missing. This can lead to overfitting [3], [26], [27] if all the instances are fully-supervised by the post-stimuli labels.

To overcome the challenge of temporal ambiguity, this paper proposes an emotion recognition algorithm based on Deep Multiple Instance Learning (*EDMIL*) using physiological signals. *EDMIL* is trained only with post-stimuli emotion labels. However, it can provide recognition results at a finer (or higher) level of granularity (every 2s) by identifying which instances represent the emotion annotated by users after watching videos. The ground truth labels (i.e., valence and arousal (V-A)) we use are based on Russell's Circumplex model [28], which describes emotions in a continuous 2-dimensional space. Valence indicates users' positive or negative affectivity. Arousal measures how calm or excited a user is. Although we use V-A for training, the prediction of *EDMIL* can be easily mapped to discrete emotion keywords (e.g., high valence and high arousal = happy, high valence and low arousal = relax) [29]. The signals and their fine-grained segments are viewed as bags and instances, respectively. Instead of implementing fully-supervised training for all the instances using post-stimuli labels, the instances are weakly-supervised by the post-stimuli labels to avoid overfitting. The fine-grained V-A of instances are then estimated by the instance gains, which represent the probability for that instance to predict the corresponding bag label. This work makes the following contributions to Affective Computing research:

- We propose an end-to-end deep multiple instance learning framework to identify which instances represent the post-stimuli V-A in a fine level of granularity using physiological signals. Our algorithm is tested on three datasets (CASE [15], MERCA [9] and CEAP-360VR [30]) collected in three environments (desktop, mobile, and HMD-based Virtual Reality (VR)). Recognition results show good performance for subject-independent 3-class (high/neutral/low) classification on all three datasets: 75.63% and 79.73% for V-A on *CASE*, 70.51% and 67.62% for V-A on *MERCA* and 65.04% and 67.05% for V-A on *CEAP-360VR*. Our framework enables finding an optimal trade-off between recognition accuracy and the burden of fine-grained emotion annotation.
- We test both state-of-the-art weakly-supervised and fully-supervised machine learning methods and compare their performance with *EDMIL*. Results show that *EDMIL*'s recognition accuracy outperforms both weakly-supervised and fully-supervised learning methods for fine-grained emotion recognition. We also find that compared with fully-supervised learning, weakly supervised learning can reduce overfitting that results from the temporal mismatch between fine-grained annotations and input signals.
- We run validation experiments to compare the performance of *EDMIL* under different instance lengths and feature extraction methods. Results show that instance segment lengths between 1-2s result in the highest recognition accuracies (up to 60% for V-A in all three datasets). Our results also show that feature extraction using an end-to-end structure can improve recognition accuracy compared with manual feature extraction and unsupervised learning feature extraction methods.

2 Related Work

In this section, we first review previous works on emotion recognition using physiological signals. We then narrow our scope to fine-grained emotion recognition and multiple instance learning based emotion recognition.

2.1 Emotion Recognition by Physiological Signals

Emotion recognition algorithms using physiological signals as input modalities can be divided into two major categories: model-specific methods and model-free methods [29]. Model-specific methods require pre-designed hand-crafted features to classify emotions from physiological signals. In general, statistical features from the time-domain (e.g., mean, standard deviation, first differential [31], [32], [33] of the signal) and frequency-domain (e.g., mean of amplitude, mean of absolute value [34], [35], signal FFT [36]) are commonly extracted for recognition. For example, Zhao *et al.* [37] extract 223 features from 4 physiological signals to recognize the valence and arousal of users. Their algorithm, which merges information from users' personality using a hypergraph learning framework, achieves up to 70% accuracy on the *ASCERTAIN* [38] dataset. Jimenez *et al.* select 13 features from PPG (4in the time domain, 9in the frequency domain) and 14 features from GSR (all in the time domain) to recognize six basic emotions. A simple k-nearest neighbor (KNN) classifier is used by Aasim *et al.* [39] to classify valence and arousal on their newly collected MULtile SEnsorial media (MULSEmedia) dataset. Similar to the accuracy from the work of Zhao *et al.* [37], they achieve 85.18% and 76.54% accuracy for valence and arousal respectively.

Although model-specific methods have been widely used by researchers for a long time [29], they require researchers to select features based on empirical experiments [29], [40]. Thus, it is costly with respect to time and does not guarantee that selected features are optimized, which limits the generalizability of their algorithms.

The model-free methods use artificial neural networks to learn the inherent structure between input signals and emotion labels. Thus, they can automatically extract features from physiological signals for recognition. Neural networks such as convolutional neural networks (CNNs) [41], [42] and Long Short-Term Memory (LSTM) networks [12], [43] are widely used for emotion recognition and achieve high accuracy. For example, a regularized deep fusion framework is designed by Zhang *et al.* [44] to learn task-specific representations for each physiological signal. Their experiments show that the method can improve the performance of subject-independent emotion recognition by 6% compared to other fusion methods such as single modal classifiers (i.e., SVM, Decision Tree, and Naive Bayes). However, model-free methods can easily overfit on the training data when using deep and sophisticated structures [45]. According to the experiments of Zhang *et al.* [3], for the one-dimensional CNN, simply deepening the network does not result in better performance on the testing set. Thus, there are still challenges in designing model-free methods for better generalizability for emotion recognition.

Our method takes advantage of the model-free methods, which automatically extract features from physiological signals. Instead of fully-supervised training, we design a weakly-supervised learning algorithm to overcome the overfitting problem of model-free methods.

2.2 Fine-Grained Emotion Recognition

Fine-grained emotion recognition requires algorithms to predict multiple emotion states by relying on signals within a specific time interval. This is typically done using two kinds of methods: regression and classification. Regression methods view the target emotion states as a continuous sequence and directly calculate the mapping (regression) from input signals to output emotion sequences. These methods include sequential learning approaches such as Long Short Term Memory (LSTM) networks [12], support vector regression (SVR) [46] and polynomial regression [47]. Classification methods on the other hand first divide the entire signals into multiple segmentations (instances) and classify the emotion for each fine-grained instance. For example, Awais *et al.* [48] designed an LSTM-based classification method to classify emotions every 5 seconds. Srinivasan *et al.* [49] implemented *decision trees* on a RaspberryPi device to classify the valence (positive/negative) of users every 10 seconds. Both the regression and classification methods need fine-grained emotion labels to train the recognition algorithm. For classification methods, the frequency of required ground truth labels is the same as the frequency of the classification results (e.g., 5s and 10s for [48] and [49], respectively). For regression methods, the frequency of required labels is usually the same as the input signals [10], [50].

To collect such fine-grained emotion ground truth labels for training, previous works either let the users themselves or professional annotators to annotate emotions at a fine granularity level. Some researchers developed momentary emotion annotation tools, such as *FEELTrace* [51], *CASE* [52], *RCEA* [9] and *RCEA-360VR* [30] which allow users to input their emotions (e.g., valence and arousal) in real-time. However, momentary annotation requires users to multi-task (e.g., watch videos and annotate at the same time), which poses risks in increasing user mental workload [9], [53] and is not always feasible if users watch longer videos [19]. For external annotators, it usually requires at least three external annotators to get a meaningful agreement (e.g., high kappa score) [16], and this requires extra labeling effort. In addition, hiring professional annotators can bring extra costs for the data collection experiment. Such challenges of collecting fine-grained emotion labels lead to researchers developing datasets such as DEAP [22], Mahnob-HCI [23] and ASCERTAIN [38] containing only post-stimuli labels. Taking advantage of these datasets can lower the cost of developing and training fine-grained emotion recognition algorithms.

In our work, we propose a fine-grained emotion recognition algorithm that is trained using only the post-stimuli emotion labels. Our method enables researchers to build a fine-grained emotion recognition model without collecting a large amount of continuously annotated emotion ground truth labels to train the recognition network.

2.3 Multiple Instance Learning Based Emotion Recognition

In the paradigm of Multiple Instance Learning (MIL), the input is a set of *bags* which are composed of multiple *instances*. At the training stage, each bag has a corresponding label while each instance does not. Thus, not all the instances are labeled the same as the bag label at the training stage [54], [55]. MIL has been applied in previous works on emotion recognition using a variety of data modalities such as images [56], text [57], voice [58] and physiological signals [19]. For physiological signals, Romeo *et al.* [19] evaluated four MIL algorithms (mi-SVM [59], mil-Boost [60], MI-SVM [59] and EMDD-SVM [61]) for emotion recognition using physiological signals (without EEG) on DEAP [22] and Consumer [19] dataset. The two datasets are collected using golden-standard (for DEAP) and unobtrusive consumer devices (for Consumer). Their results show that mi-SVM and MI-SVM achieve the highest recognition accuracies (bag level) on DEAP dataset, which is 63.6% and 61.1% for valence and arousal, respectively. The hypothesis of the four methods mentioned above is that the positive bags are fairly rich in positive instances. Thus, the negative instances can be easily identified. However, positive bags can contain only a small fraction of positive instances. To solve this problem, Bunescu *et al.* [62] designed Balanced MIL (sb-MIL) which introduced a balancing constraint between positive and negative instances to model the sparse positive instances in different bags. Zhang *et al.* [63] implemented the proposed sb-MIL [62] to classify dimensional emotions using EEG signals. They achieved classification accuracies (bag level) on DEAP [22] dataset of 74.21% and 77.50% for valence and arousal, respectively.

The general idea of the MIL algorithms mentioned above is to identify the instances which contribute to maximize the probability for predicting the bag labels. However, all these methods need to manually design loss functions or

constraints between instances and bags [59], [62]. The manually designed functions or constraints are usually just suitable for one condition (e.g., most of the instances are labeled the same as the bag label [59]). In addition, the previous works mentioned above only recognize emotions at the bag level, which means they recognize only one emotion instead of the fine-grained emotion response for each instance (i.e., instance-level recognition). In the work of Roman *et al.* [19], the authors attempt to identify the instances which make contributions to predicting the bag-level labels. However, since the two datasets they use do not have fine-grained emotion ground truth labels, they plan to validate their method for fine-grained emotion in future work.

Compared with other learning tasks, the special character of fine-grained emotion recognition using physiological signals is manifested into two aspects. First of all, most of the existing multi-instance learning methods [56], [64], [65] for emotion recognition are designed for 2D images. The purpose of these methods is to identify an object of interest embedded into the image (also known as "*emotional regions which contain objects and concepts*" according to the definition of Zhao *et al.* [66]). Thus, the instances (i.e., small segments of the image) are often aggregated in a dense spatial space. Strong constraint functions are often implemented to omit instances which are spatially far away from the region with the highest probability of an object of interest. For example, in the work of Rao *et al.* [56], a linear iterative clustering is used to merge different regions of interest representing the emotion of images. Compared with spatial differences of emotions from an image, the dynamics of emotions are sparsely distributed in the temporal space: users can have an emotion response with a short duration (0.5s to 2s) [6], [7]. Thus, we did not implement strong constraints to filter the instances which have high probability of predicting the emotions but are not densely aggregated in one temporal moment. We use a simple threshold based on the distribution of the instance gains for identifying which instances correspond to the post-stimuli emotion labels.

Secondly, compared with learning tasks using other temporal signals (e.g., video, speech, EEG signals), the physiological signals we use contain less abundant information for emotion recognition [19], [67]. This limits the performance of DNN methods: feature extraction layers with deep structures can easily overfit or fail to extract meaningful features for recognition [3]. Due to this, we use shallow convolutional layers (5-layers) with gradually increasing number of filters and lower size kernels for feature extraction. This character also motivates us to compare two different types of feature extraction methods in section 5.6 to find out whether the end-to-end deep-network-based feature extraction is suitable for fine-grained emotion recognition using physiological signals.

3 Deep MIL Based Emotion Recognition

In this section, we propose a deep multiple instance learning based emotion recognition algorithm (*EDMIL*) to identify the post-stimuli dimensional emotions (i.e., valence and arousal (V-A)) at a fine granularity level from physiological signals. *EDMIL* recognizes fine-grained V-A by identifying which instances represent the post-stimuli emotion labels annotated by users after watching the videos. In the training stage, *EDMIL* contains four parts: (1) *Pre-processing:* the obtained physiological signals are firstly filtered and grouped into bags and instances as input for *EDMIL*. (2) *Feature extraction layers*: the grouped signals are then passed into deep convolutional layers for feature extraction. (3) *Multiple instance learning layers*: the extracted features are then input into multiple instance learning layers to obtain the instance gain for each instance. The instance gain represents the probability for each instance to predict the bag label. (4) Fully connected layer: at last, each instance gain is fully connected with the post-stimuli emotion labels (i.e., valence or arousal). The training network is designed to learn the data representation to predict the post-stimuli emotion labels using the entire signals. The instance gains learned in part (3) are the matching scores which indicate the probability that the instance contributes to the prediction of the post-stimuli emotion label. In the prediction stage (the network has already been trained and fixed), the obtained signals are first forwarded from (1) to (3) to get the instance gains. After that, the (5) instance regularization is used to transfer the instance gains into the V-A for each instance. The architecture of the algorithm is shown in Fig. 1. When describing and validating *EDMIL*, we use the physiological signals as input and specify the application scenario as video watching.

3.1 Pre-Processing

We first pre-process all the physiological signals using different filters to eliminate the noise and artifacts from the measurement. The details for this process are described in section 5.1 (implementation details). Suppose $S_{mn} = \{s_c\}_{c=1}^{C}$ is the set of pre-processed physiological signals for one user m watching one entire video n, where C is the number (channels) of physiological signals. The signals are firstly segmented into multiple instances with a fixed instance L. After the segmentation, the input of the algorithm is transferred into a bag of instances: $B = \{b_g\}_{g=1}^{G}$. G is the number of samples for training. $b_g = \{x_g^i\}_{i=1}^{I}$ is the bag g and x_g^i is the instance i in bag g. $b_g \in R^{L\times I\times C}, x_g^i \in R^{I\times C}$, where L and I is the number of instances in one bag and the length for one instance, respectively. The goal of *EDMIL* is to predict the V-A for each instance. For both the training and prediction stage, only the ground truth labels for b_g are available. The ground truth labels for x_g^i are only used to evaluate the performance of *EDMIL*.

3.2 Feature Extraction Layers

The feature extraction layers are designed to learn the deep features from the physiological signals for recognizing the post-stimuli labels. The features are extracted from each instance x_g^i independently, which means the feature extraction layers will not influence the independence between each set of instances (no features are extracted from multiple instances). This operation guarantees that each instance has a unique instance gain before the fully connected layer. Here, three types of feature extraction methods are implemented for comparison: (1) an end-to-end feature extraction method using one-dimensional convolutional neural network (1D-CNN) (*deepfeat*), (2) an unsupervised feature extraction method by maximizing the correlation coefficients between pairwise physiological signals (*pcorrfeat*) and

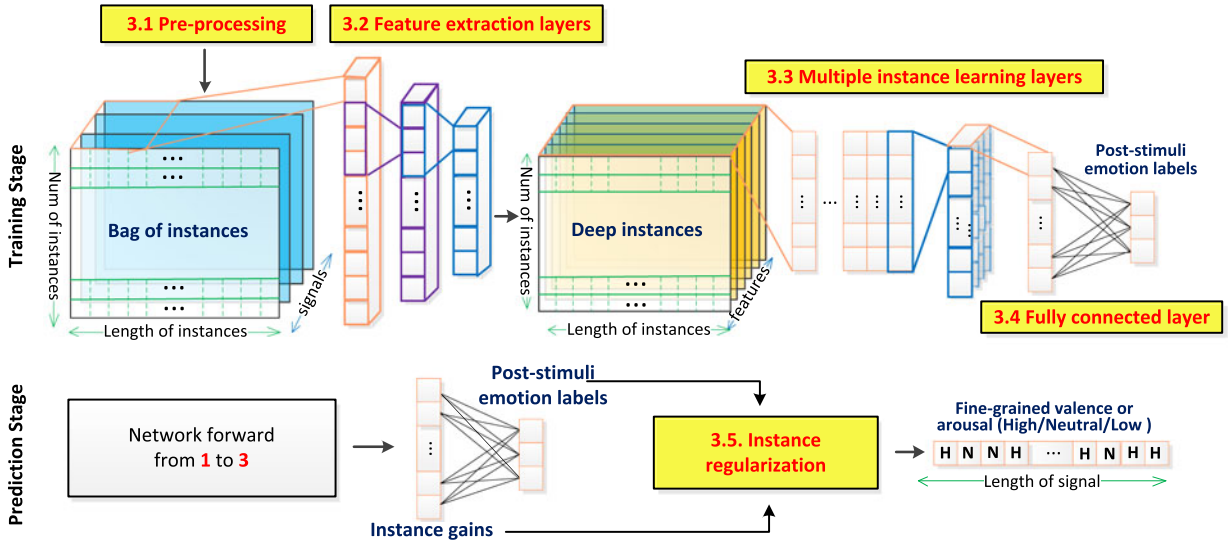

Fig. 1. The architecture of proposed EDMIL.

(3) a manual feature extraction method using statistical features (*manualfeat*). Unless otherwise specified, we use *deepfeat* as the default feature extraction method.

Theoretically, the end-to-end model should result in the best performance as the features are directly connected with the ground truth labels [68], which means the deep representation is trained to best recognize these labels. However, according to previous studies [3], [69], if we train the network using fine-grained emotion labels and fully-supervised learning methods, the end-to-end model will overfit because of the temporal resolution mismatch between physiological signals and fine-grained self-reports due to different interoception levels across individuals [70]. Thus, we compare these three types of methods to find out whether the end-to-end, deep feature extraction (*deepfeat*, section 3.2.1) still has the problem of overfitting for weakly-supervised learning. Manual feature extraction methods (*manualfeat*, section 3.2.3) are widely used by previous works for emotion recognition [29], [40], [71]. Thus, we choose it as a baseline method for comparison. We additionally compare *deepfeat* with an unsupervised feature extraction method (*pcorrfeat*, section 3.2.2) because it can decrease overfitting compared with end-to-end models, as shown in prior work by Zhang *et al.* [3]. Below, we introduce the details of the three feature extraction methods.

3.2.1 Deepfeat

The deep features (*deepfeat*) are extracted using a 5-layer 1D-CNN [72]. The parameters for each convolutional layer are shown in TABLE 1. We use large (i.e., equals to half of the instance length) convolutional kernels in the beginning of the network. Large convolutional kernels commonly result in better recognition accuracy [73] because they have a large receptive field across different sampling points in one instance. However, large kernels can omit the local information and make the network more difficult to converge [74]. Thus, we follow a classical strategy that gradually increases the number of kernels and decreases the size of them when the network goes deeper [75], [76]. At last, we add a convolutional layer whose size is bigger than the previous layer to merge the local information learned by small kernels.

After the feature extraction layer, the bag of instances B is transferred into deep instances $D = \{d_g\}_{g=1}^{G}, d_g = \{f_g^k\}_{k=1}^{K}$ where $K = 128$ is the dimension of features at the last 1D-CNN layer.

3.2.2 Pcorrfeat

The pairwise correlation-based features (*pcorrfeat*) are extracted by maximizing the correlation coefficient for every two signals from users who watch the same video stimuli [69]. The idea is inspired by the hypothesis that the same stimuli will trigger relatively similar emotions across physiological responses among different users [77], [78]. To extract correlation-based features, we first calculate the covariance (C_{11} and C_{22}) and cross-covariance (C_{12}) of the two signals (S_n^1, S_n^2) for users who watch the same video stimuli. After that, we implement the Singular Value Decomposition (SVD) on the equation below:

$$[U, D, V] = \mathrm{SVD}(V_1 D_1 V_1^T \cdot C_{12} \cdot V_2 D_2 V_2^T) \tag{1}$$

where D_1 and D_2 are diagonal matrices whose diagonal elements are the ω biggest non-zero eigenvalues of C_{11} and C_{22}, respectively. $D_1 = \mathrm{diag}(\frac{1}{\sqrt{D_{11}}}, \frac{1}{\sqrt{D_{12}}}, \ldots, \frac{1}{\sqrt{D_{1\omega}}})$ and D_2 have the same format. $V_1 = [V_{11}, V_{12}, \ldots, V_{1\omega}]$ is composed of the ω corresponding eigenvectors of $[D_{11}, D_{12}, \ldots, D_{1\omega}]$,

TABLE 1
Architecture of the 1D-CNN to Extract *Deepfeat*

layer	input size	channels	kernel size	output size
input	(I,C)	8	I/2+1*	(I,8)
conv1	(I,8)	16	I/3	(I,16)
conv2	(I,16)	32	I/4+1*	(I,32)
conv3	(I,32)	64	I/8+1*	(I,64)
conv4	(I,64)	128	I/12+1*	(I,128)
conv5	(I,128)	128	I/8+1*	(I,128)

*We add 1 to some of the kernels to make their size an odd.

respectively. V_2 is calculated using the same method. We then obtain two linear projections $[H_1^t, H_2^t] = [V_1 D_1 V_1^T \cdot U,' V_2 D_2 V_2^T \cdot V']$, where U' and V' consist of the first K columns of U and V, respectively. At last, the correlation-based features of S_1^t and S_2^t can be obtained by: $F^t = [S_1^t \cdot H_1^t, S_2^t \cdot H_2^t]$. We then implement the above procedure among all the M stimuli and C signals (pair by pair). At last, the bag of instances B is transferred into *pcorrfeat* $P = \{p_g\}_{g=1}^{G}, p_g = \{f_g^k\}_{k=1}^{K}$ where $K = \omega \prod_{i=2}^{C-1} i$ is the dimension of *pcorrfeat*.

3.2.3 Manualfeat

For the manually selected features, we select the features both in the time and frequency domain. These are widely used features for physiological signals for the baseline comparison in dataset and review papers [29], [40], [71] for affective computing. For the features in the time domain, we choose the mean, standard variance, average root mean square, mean of the absolute values, maximum amplitude and average amplitude for the original and first-order differential of all physiological signals. For the features in the frequency domain, we choose the mean, maximum and the magnitude for the Fast Fourier Transform (FFT) [79] of all physiological signals.

3.3 Multiple Instance Learning Layers

The purpose of the multiple instance learning layers is to model the probability between an instance and the corresponding post-stimuli V-A labels [55]. According to the *peak-end theory* [25], the post-stimuli labels usually represent the most salient (peak) or recent (end) V-A within the entire video watching rather than the fine-grained V-A changes. Thus, only a part of the instances inside one bag represent the post-stimuli emotion labels. Traditional MIL algorithms have to make a hypothesis that instances corresponding to the bag label are densely [59] or sparsely [62] consist of the bag. Unlike traditional MIL algorithms, we designed two multiple instance learning layers to automatically learn the instance gains without a pre-set hypothesis. The multiple instance learning layers assign each instance a matching score (instance gain) which maximize the probability to predict the post-stimuli V-A using the whole bag. That means instances which can better enable the network to predict the post-stimuli V-A will be assigned higher instance gains.

The diagram for multiple instance learning layers is shown in Fig. 2. For each bag $d_g \in R^{L \times I \times K}$, a maximum pooling is implemented at the feature level (at dimension of features K) to select the biggest features for each time point i. After that, we activate the features f_g^k for instance k as:

$$a_{k,g} = \Psi(\alpha_{k,g} f_g^k + \beta_{k,g}) \tag{2}$$

$\alpha_{k,g}$ and $\beta_{k,g}$ are the weight and bias for the activation operation respectively. $\Psi(\cdot)$ represents the activation function. Here, we use a *softmax* function according to previous works [55], [80]:

$$\Psi(i) = \frac{e^i}{\Sigma_j e^j} \tag{3}$$

The purpose of the activation operation is to (a) normalize the selected features in the range from 0 to 1 and (b) make it easier for the network to calculate the gradient during back-propagation. At last, another max pooling operation is implemented at dimension of instance length I. After that, we obtain the instance gains $Z = \{z_g\}_{g=1}^{G}, z_g \in R^{1 \times L}$ with the same dimension of the number of instances L.

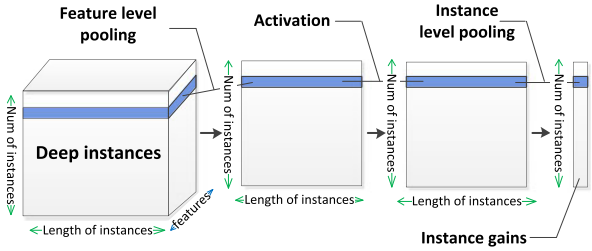

Fig. 2. The diagram for multiple instance layers.

3.4 Fully-Connected Layer

To build the link between the instance gains and post-stimuli emotion labels, we put one fully connected layer at the end of *EDMIL*. For the multi-class (high/neutral/low V-A) classification task, we use the *softmax* in (3) as the activation function. Then we train the network using *RMSprop* [81] optimizer because *RMSprop* can automatically adjust the learning rate for faster convergence. Since the task is multi-class classification, we use the categorical cross entropy (H_c) as the loss function for training:

$$H_c = -\frac{1}{n} \Sigma_i [y_i \cdot \ln x_i + (1 - y_i) \cdot \ln(x_i)] \tag{4}$$

where x and y are the predicted and true value for the fully-connected layer, respectively.

The target of the network is to learn the data representation to predict the post-stimuli emotion labels using the signals for the whole video watching. Thus, the training for post-stimuli labels is fully-supervised. However, the information for which instances can represent the emotion users labeled post-stimuli is not available during the training stage. The instance gain is only the probability of whether the instance makes contributions for the bag to predict the post-stimuli labels. Thus, for each instance, the training is weakly-supervised.

3.5 Instance Regularization

In the prediction stage, when a new user watches a video, their physiological signals are forwarded from pre-processing (section 3.1) to multiple instance learning layers (section 3.3) to get the instance gains. After that, the instance regularization is implemented to identify which fine-grained instances are correlated with the post-stimuli V-A. Since the instance gain only represents the matching score, we need to obtain the post-stimuli label the user annotated to know which V-A value these instances match. Thus, the post-stimuli V-A is also needed in this step, which means the user needs to input his or her V-A after watching the entire video. We predict the fine-grained V-A according to both the instance gain and the annotated V-A after watching the video:

$$y_i = \begin{cases} Y, & z_i > mean(Z) \\ p, & z_i <= mean(Z) \end{cases} \tag{5}$$

Fig. 3. The experimental setup and annotation interface (©[2020] IEEE) for CASE [15] .

where $Z = \{z_i\}$ is the instance gains. mean(z) is the mean value of all instance gains in one bag (signals for the user watch the entire video). y_i is the predicted V-A for instance i. Y is the post-stimuli V-A annotated by the user. p is the baseline V-A (i.e., neutral).

4 Datasets

To evaluate the performance of *EDMIL*, we test it on three datasets: *CASE* [15], *MERCA* [9] and *CEAP-360VR* [82] collected in three environments: desktop, mobile and HMD-based Virtual Reality, respectively. *EDMIL* is an end-to-end weakly-supervised learning algorithm for modeling the temporal ambiguity of emotions. Thus, we choose physiological signals as the uni-dimensional input for testing the validity of *EDMIL* according to previous work on MIL based emotion recognition [19].

All the three datasets we choose contain physiological signals with fine-grained self-reported valence and arousal. The fine-grained self-reports are used for validating the accuracy of *EDMIL*. In the training and prediction stage, *EDMIL* does not need fine-grained self-reports as input information. We evaluate *EDMIL* on datasets collected in three different environments to test whether it can be generalized to different application scenarios. In addition, the signals are collected using both golden standard and wearable devices. Evaluating *EDMIL* on these three datasets can also test whether it can generalize to different types of physiological sensors. The details of datasets are described below.

4.1 CASE Dataset

The *CASE* (Continuously Annotated Signals of Emotion) dataset [15] contains physiological signals from 30 participants (15m, 15f), aged between 22-37 (M=27.1, SD=3.9). Participants used a physical joystick (shown in Fig. 3) to annotate their valence and arousal continuously while they watched eight video clips on a desktop screen. The data collection experiment for CASE is a 1 (task: watch videos and continuously annotate valence and arousal) × 4 (video emotions: amusing vs. boring vs. relaxing vs. scary) within-subject design. Eight video clips (two videos per emotions, duration M=158.75s and SD = 23.67s) selected from movies and documentaries were chosen to elicit the 4 emotions. The experiment was conducted in an indoor laboratory environment. Six clinical, golden standard physiological sensors (Electrocardiogram (ECG), Blood Volume Pulse (BVP), Electrodermal activity (EDA), Respiration (RESP), Skin Temperature (TEMP), Electromyography (EMG)) were equipped to collect physiological signals. All sensors were synchronized using a specialized hardware and sampled at 1000Hz (sample size = 2451650 samples × 30 users). The V-A ratings (sample size = 49033 samples × 30 users) were collected at 20Hz according to the sampling rate of the physical joystick.

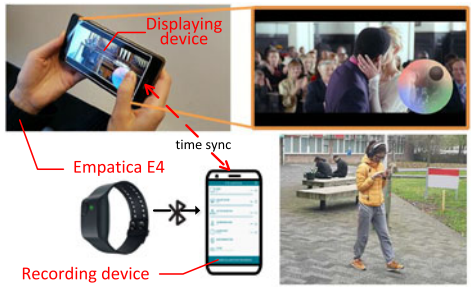

Fig. 4. The experimental setup and annotation interface for MERCA [9].

4.2 MERCA Dataset

The *MERCA* [9] (Mobile Emotion Recognition dataset with Continuous Annotations) dataset contains physiological signals for 20 participants (12m, 8f) aged between 22-32 (M = 26.7, SD = 2.9). Users used a virtual joystick (shown in Fig. 4) to annotate their valence and arousal continuously while watching 12 video clips on a mobile screen (5.5 in). The data collection experiment for MERCA is a 1 (task: watch videos and continuously annotate valence and arousal) × 4 (video emotions: joy vs. fear vs. sad vs. neutral) within-subject design. The 12 video clips (three videos per emotion, duration M = 81.4s and SD = 22.5s) were selected according to 2D emotion annotations from the self-reports in the MAHNOB-HCI dataset [23]. 10s black screens were added before and after each video to decrease the effect of emotional overlapping among different videos.

As shown in Fig. 4, the experiment was conducted in an outdoor campus. Users could walk or stand freely while watching videos. The experiment setting parallels watching mobile videos while walking or waiting for a bus or train, which is a common phenomenon in mobile video consumption [83], [84], [85]. Participants were told to watch the videos as they normally would in such settings. To prevent participants from running into obstacles, traffic, or other people, the experimenter always accompanied the participant from a distance to guarantee their safety.

The Empatica E4[1] wristband was used to collect physiological signals. Empatica E4 is a non-intrusive and wearable wristband which is suitable for collecting signals in outside environments. From Empatica E4, they collected Heart Rate (HR, 1Hz, sample size: 1326 samples × 20 users), BVP (64Hz, 42432 samples × 20 users), EDA (4Hz, 5304 samples × 20 users) and TEMP (4Hz, 5304 samples × 20 users). The collected signals were stored on a mobile device (i.e., the recording device). As shown in Fig. 4, the E4 wristband were connected to the recording device through low-power bluetooth. Another mobile device (i.e., the displaying device) was used for showing the videos and collecting annotations. MERCA also contains eye movements from a wearable eyetracker. A noise-cancelling headphone was connected to the displaying device via bluetooth to play audio. Timestamps of both devices were set according to the clock of the recording device, where all data were

1. https://www.empatica.com/en-eu/research/e4/

Fig. 5. The experimental setup and annotation interface for CEAP-360VR [82].

synchronized via an NTP server (android.pool.ntp.org). The V-A ratings (sample size = 13260 samples × 20 users) were collected at 10Hz according to the sampling rate of the virtual joystick. At the end of video watching, users were asked to rate their post-stimuli V-A for that video using the Self-Assessment Manikin (SAM) scale [86].

4.3 CEAP-360VR Dataset

The *CEAP-360VR* [82] (Continuous Physiological and Behavioral Emotion Annotation Dataset for 360° Videos) dataset contains physiological signals for 32 participants (16m, 16f) aged between 18-33 (M = 25, SD = 4.0). Similar to CASE and MERCA, a physical joystick (Joy-Con Controller, shown in Fig. 5) was used by users to annotate their valence and arousal continuously while they were watching eight 360° video clips through an HTC Vive Pro Eye[2] Head-Mounted Display (HMD). The data collection experiment is a 1 (task: watching 360° videos and continuously annotate valence and arousal) × 8 (video emotions: high valence+high arousal vs. high valence+low arousal vs. low valence+low arousal vs. low valence+high arousal) within-subjects design. The eight video clips (two videos per emotions, duration = 60s) were selected according from the database provided by Li *et al.* [87], which contains mean post-stimuli V-A ratings from 95 subjects.

As shown in Fig. 5, the experiment was conducted in a controlled, indoor environment. During the experiment, participants sat on a swivel chair and were free to look in any direction [30]. The experimental setup parallels the scenario that users watch 360° videos using HMD-based VR devices. Similar to MERCA, the physiological signals of participants were measured through the Empatica E4 wristband. For Empatica E4, they collect HR (1Hz, sample size: 360 samples × 32 users), BVP (64Hz, 11520 samples × 32 users), EDA (4Hz, 1440 samples × 32 users) and TEMP (4Hz, 1440 samples × 32 users). The collected signals were stored on a mobile device which was connected with E4 using low-power bluetooth. One laptop was used to play the 360° videos as well as log the head and eye movement from HTC Vive Pro Eye. Timestamps of the mobile device were set according to the clock of the the laptop, synchronized via an NTP server. The V-A ratings (sample size = 3600 samples × 32 users) were collected at 10Hz according to the sampling rate of the physical joystick. After each 360° video was played, users were ask to rate their post-stimuli V-A for that video using a within-VR SAM [86] rating scale.

5 Experiments and Results

In this section, we first introduce the implementation details of *EDMIL* on CASE, MERCA and CEAP-360VR datasets. We then evaluate the classification and regression performance of *EDMIL* using Leave-One-Subject-Out Cross Validation (LOSOCV). After that, we compare the performance of *EDMIL* with the state-of-the-art MIL algorithms which have been applied for emotion recognition. Then, an ablation study was conducted to verify the effectiveness of each component. Lastly, we compare the performance of the three feature extraction methods mentioned in section 3.2.

5.1 Implementation Details

For all three datasets, we choose four physiological signals, Electrodermal activity (EDA), Blood Volume Pulse (BVP), Skin Temperature (TEMP) and Heart Rate (HR), as the input signals of *EDMIL*. Although EEG signals can provide more abundant information according to previous works [29], [88], high-resolution EEG signals need to be captured under strict laboratory environments without any electromagnetic interference [89], which makes their use limited in an indoor laboratory environment. We choose these four signals because they can be easily measured by wearable and unobtrusive sensing devices such as smart watches or wristband (e.g., Empatica E4 and Microsoft MS Band[3]). In addition, the selected signals contain physiological responses from both autonomic nervous system (EDA and TEMP) and cardiovascular system (BVP and HR), which can provide abundant information for emotion recognition [29], [88]. Moreover, these four signals have also been widely used by previous work to recognize valence and arousal [90], [91], [92]. We first pre-process the physiological signals using the standard filtering procedure widely used in previous works [29], [93], [94], [95]. Firstly, a low pass filter with a 2Hz cutoff frequency is used to remove noise [93] from EDA signals. For the BVP signal, we implement a 4-order butterworth bandpass filter with cutoff frequencies [30, 200] Hz to eliminate the bursts [94]. At last, an elliptic band-pass filter with cutoff frequencies [0.005, 0.1] is used to filter TEMP signals [95].

To decrease measurement bias in different sessions, all signals are normalized to [0,1] using Min-Max scaling normalization:

$$S_n = \frac{S - \min(S)}{\max(S) - \min(S)} \tag{6}$$

The normalization is implemented on each subject under each video stimulus (session). Since signals in both MERCA and CEAP-360VR have different sampling rates, they are interpolated to 50Hz using linear interpretation [96]. We choose linear interpolation because it is the simplest interpolation method which will not change the distribution of the signals. For CASE dataset, we downsample the signals also to 50Hz by decimation down-sampling [97]. The HR for CASE is extracted from ECG using *heartpy* library [98]. Then the input signals are segmented into 2 s instances (sample size 100). The choices for different segmentation lengths are discussed in Section 6.2. Since different sessions have different lengths, we use zero padding according to previous works [55], [99] to let all sessions (i.e., bags) have the same length. Since CASE does not collect post-stimuli

2. https://enterprise.vive.com/us/product/vive-pro-eye/ 3. http://developer.microsoftband.com

V-A, we use the mean of continuous V-A as ground truth to train *EDMIL* because the mean V-A has no significant difference between post-stimuli V-A [9], [30]. Aside from the comparison of different feature extraction methods (section 5.6), we use *deepfeat* described in section 3.2.1 for feature extraction. The network is trained by *RMSprop* [81] optimizer since it can automatically adjust the learning rate for faster convergence. We use the *Early-Stopping* [100] technique to terminate training if there is no improvement on the training loss for 5 epochs.

The time complexity of *EDMIL* is:

$$O\left(\frac{15}{64}K^2 \cdot I^2 \cdot L^2 + \frac{1}{32}K \cdot C \cdot I^2 \cdot L + \left(\frac{69K^2}{32} + \frac{K \cdot C}{16} + 2\right) \cdot I \cdot L + I\right) \tag{7}$$

L is the number of instances in one bag. I is the number of sample points for one instance. $N = L \times I$ is the sample size of an input signal. K and C are constants which represent the output dimension of the feature vectors and the number of signal channels respectively. Thus, the time complexity can be simplified as:

$$O(A \cdot N^2 + B \cdot N + D) \tag{8}$$

where A, B are coefficients for N^2 and N respectively. D is the constant term of the time complexity. The average training time of *EDMIL* is 218.56s, 156.39s and 192.23s for CASE, MERCA and CEAP-360VR, respectively. *EDMIL* is implemented using Keras (python). All our experiments are run on a server with NVIDIA RTX 2080Ti GPU and 32GB RAM. The average testing time for each fine-grained instance is 19.21ms, 18.6ms and 15.34ms for CASE, MERCA and CEAP-360VR, respectively. That means to recognize 2s emotions, the algorithm only spends less than 20ms after the network is trained. The computational cost of *EDMIL* is low due to (a) the simple (5-layer) structure for the feature extraction (b) a simple threshold instead of complex constraint functions for the instance regularization module.

5.2 Evaluation Protocol

To evaluate the performance of *EDMIL*, we conduct two kinds of experiments: classification and regression. The classification task tests the instance-level accuracy while the regression task validates the overall dynamics throughout one entire video watching. Below, we introduce the details of the two tasks as well as the metrics and method we use to validate them.

5.2.1 Classification Task

The aim of the classification task is to test whether *EDMIL* can recognize high/neutral/low V-A for each instance, which is a standard validation method in prior works [23], [29], [101]. We use the mean V-A of instances as ground truth labels for validation. The mapping from continuous values of V-A to discretized categories is: [1,3) = Low, [3, 6) = Neutral, [6, 9] = High.

To test the performance of the classification task of *EDMIL*, we select three validation metrics:

- *accuracy (acc):* the percentage of correct predictions;
- *confusion matrix:* the square matrix that shows the type of error in a supervised paradigm [19];
- *weighted F1-score (w-f1):* the harmonic mean of precision and recall for each label (weighted average by the number of true instances for each label) [102].

These three metrics are widely used in evaluating machine learning algorithms [103]. We use weighted F1-score instead of macro and binary F1-score to take into account label imbalance.

5.2.2 Regression Task

The performance of the classification task can only reflect the pairwise comparison between the predicted and ground truth labels for instances, not the overall difference between sequences (i.e., the predicted and ground truth V-A throughout one entire video watching). The purpose of the regression task is to test whether the instance gains (before instance regularization) learned from post-stimuli V-A have similar temporal dynamics with fine-grained V-A ground truth. In addition, as a discretization step, the instance regularization could bring bias to the classification performance. The predicted and ground truth V-A may be different but be discretized into the same category. Thus, the test of regression task can provide an additional validation of *EDMIL*.

To compare the performance of regression, we also train *EDMIL* with 3-class (low/neutral/high) post-stimuli V-A labels to get the instance gains. We then skip the instance regularization and compare the obtained instance gains and fine-grained V-A labels. Since the instance gains and V-A have different magnitudes, we also normalize them using min-max scaling normalization. To evaluate their difference, we use the mean square error (mse) as the validation metric for the regression task:

$$mse = \frac{1}{M}\Sigma_i(y_i - x_i)^2 \tag{9}$$

where y_i and x_i are the ground truth and predicted V-A for instances respectively. M is the number of instances in one bag.

5.2.3 Evaluation Method

We train and test the proposed method using subject-independent models. The subject-independent model is tested using Leave-One-Subject-Out Cross Validation (LOSOCV). LOSOCV is a standard validation method for emotion recognition which can be used to test the generalizability among different users [22]. Data from each subject are separated as testing data and the remaining data from other subjects are used for training. We repeat the training and testing operation for N times (N is the number of subjects in one dataset) to make sure the data from all subjects are used for testing. The results we show are the averaged accuracy, w-f1 and mse among all subjects used as testing data.

5.3 Results

The classification and regression performance of *EDMIL* on three datasets are shown in Table 2. The accuracies for 3-class classification for all three datasets are above 65%. The w-f1 scores are also higher or equal to 0.65, which means

TABLE 2
LOSOCV Reuslts for CASE, MERCA and CEAP-360VR

		acc	w-f1	mse
CASE	**valence**	75.63%	0.72	0.2354
	arousal	79.73%	0.77	0.2281
MERCA	**valence**	70.51%	0.69	0.2673
	arousal	67.62%	0.65	0.2051
CEAP-360VR	**valence**	65.04%	0.65	0.2384
	arousal	67.05%	0.65	0.2529

EDMIL can provide balanced recognition precision and recall for different V-A categories. Fig. 7 shows the confusion matrices for classification. For the comparison between different datasets, *EDMIL* performs the best on CASE dataset (up to 75% accuracy for V-A). The recognition accuracies on CEAP-360VR and MERCA are similar (around 67% for V-A) but lower than the accuracies on CASE. The results indicate that the mobile and VR environments are more challenging for fine-grained emotion recognition compared with a laboratory-based desktop environment. Although the performance on different datasets are different, they all achieve promising accuracies (>65%) and w-f1 scores (>0.65). The test results on different datasets show good generalizability of *EDMIL* among different testing environments (desktop, mobile and VR).

Fig. 6 shows the accuracy of 3-class classification for each individual user in the three datasets. From the results we can find variability of recognition accuracy among different individuals. The results are coherent with the work of Koelstra *et al.* [22] and Romeo *et al.* [19] that there is high inter-subject variability of physiological signals which affects the recognition accuracy. However, the accuracies for more than 75% users (CASE: 93%, MERCA: 85%, CEAP-360VR: 75% of the users respectively) are above 60%. Thus, *EDMIL* achieves balanced performance on different users, which shows good generalizability of *EDMIL* among different subjects.

5.4 Comparison With Baselines

The comparison of *EDMIL* with baseline methods [59], [62], [104], [105], [106], [107] is shown in Table 3. We choose four classic multiple instance learning algorithms which have been widely used as baseline methods. In the work of Romeo *et al.* [19], mi-SVM and MI-SVM [104] achieved the best recognition accuracies (in bag-level) for emotion recognition using physiological signals. We then add another two baseline methods, NSK [59] and sb-MIL [62], to further compare the performance of *EDMIL* with state-of-the-art methods. For these four methods, we use the same hand-crafted

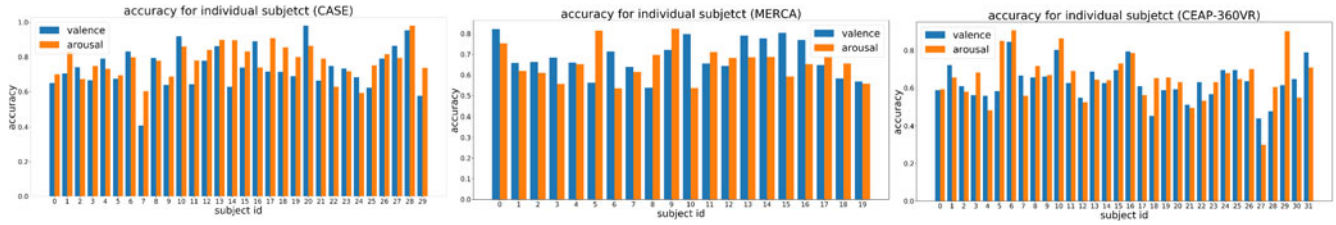

Fig. 6. The LOSOCV accuracy for individual subject of three datasets.

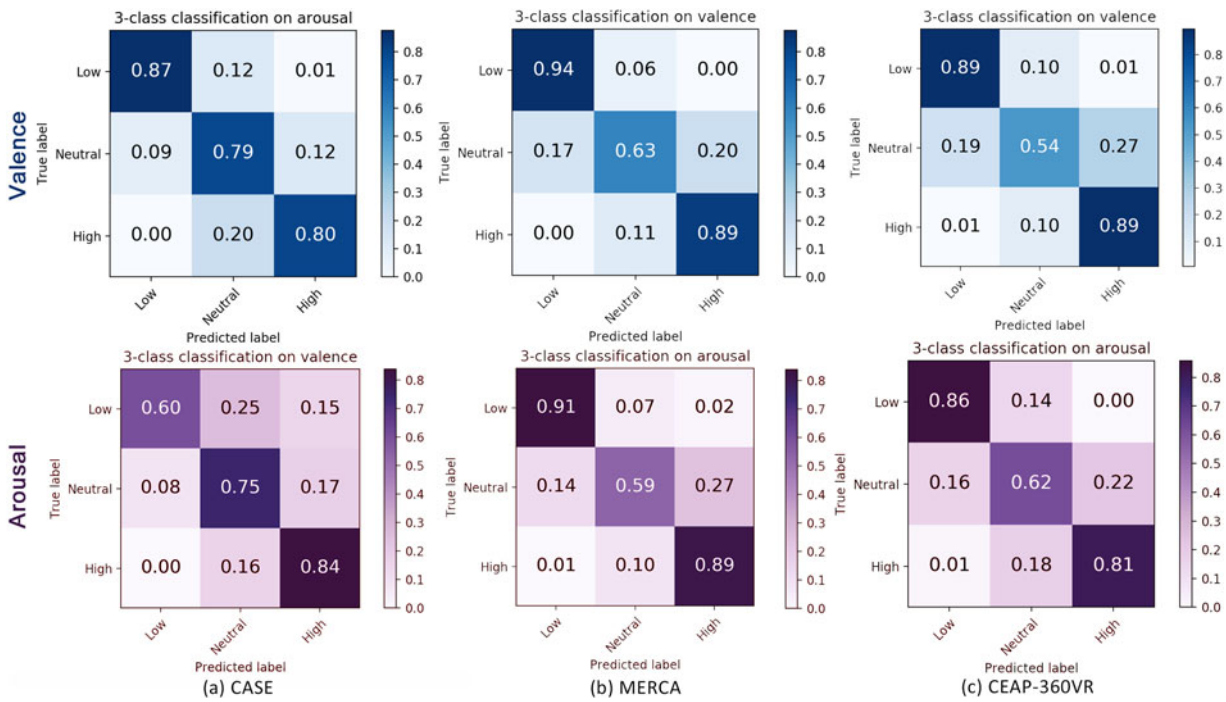

Fig. 7. The confusion matrices for leave-one-subject-out cross validation (3-class classification) on (a) CASE, (b) MERCA and (c) CEAP-360VR.

TABLE 3
Comparison With Baseline Methods

		CASE			MERCA			CEAP		
		acc	w-f1	mse	acc	w-f1	mse	acc	w-f1	mse
mi-SVM [104]	valence	53.55%	0.60	0.3856	52.41%	0.46	0.3956	49.79%	0.44	0.3845
	arousal	55.45%	0.57	0.3927	54.93%	0.49	0.4012	52.90%	0.47	0.3822
MI-SVM [104]	valence	56.45%	0.53	0.3543	52.41%	0.46	0.3726	50.21%	0.45	0.3733
	arousal	59.26%	0.55	0.3862	55.07%	0.48	0.3852	47.10%	0.41	0.3871
NSK [59]	valence	56.45%	0.54	0.3774	47.14%	0.46	0.3845	48.09%	0.45	0.4151
	arousal	59.66%	0.55	0.3983	55.07%	0.48	0.4011	49.10%	0.51	0.4169
sb-MIL [62]	valence	57.47%	0.53	0.2873	52.41%	0.46	0.2991	50.21%	0.49	0.2927
	arousal	58.66%	0.55	0.2911	47.07%	0.50	0.3012	47.00%	0.47	0.2913
AMIL [105]	valence	66.45%	0.64	0.2451	59.59%	0.51	0.2745	55.87%	0.45	0.2457
	arousal	68.57%	0.62	0.2326	58.32%	0.48	0.2677	57.62%	0.48	0.2678
WSPPG [106]	valence	70.26%	0.61	0.2382	65.34%	0.53	0.2734	61.54%	0.51	0.2403
	arousal	71.32%	0.62	0.2387	62.51%	0.54	0.2232	63.18%	0.53	0.2612
WCRNN [107]	valence	61.43%	0.51	0.2874	55.11%	0.43	0.3052	49.19%	0.41	0.3037
	arousal	63.37%	0.53	0.2943	50.67%	0.41	0.2873	50.13%	0.45	0.3315
EDMIL	valence	**75.63%**	**0.72**	**0.2354**	**70.51%**	**0.69**	0.2673	**65.04%**	**0.65**	**0.2384**
	arousal	**79.73%**	**0.77**	**0.2281**	**67.62%**	**0.65**	**0.2051**	**67.05%**	**0.65**	**0.2529**

features with [19]. In addition to the classic MIL algorithms, we also select three deep learning based weakly-supervised methods for comparison (i.e., Attentional Multiple Instance Learning (AMIL) [105], Weakly Supervised PPG (WSPPG) [106], Weakly supervised Convolutional Recurrent Neural Network (WCRNN) [107]). The baselines we choose include some widely used and more complex machine learning structures (i.e., recurrent structure [107], attention structure [105], deeper CNN [105], [105]). All these three baselines use the end-to-end learning structures which are designed for 1D signal based (voice [105], [107], PPG [106]) learning tasks. Thus, we can directly use them for testing without manually selecting features like classic MIL baselines. Similar to *EDMIL*, we use these seven methods to obtain the instance gains to compare the regression performance. For the classification performance, we use the same instance regularization to transfer the instance gains into fine-grained V-A for all the four baseline methods.

As shown in Table 3, the performance of *EDMIL* outperforms all four classic MIL baseline methods. The results are coherent with the finding of Romeo *et al.* [19] that classic MIL algorithms cannot achieve high recognition accuracy using subject-independent models. The classic MIL algorithms need to make hypotheses that the instances corresponding to the bag label are densely (mi-SVM, MI-SVM and NSK) or sparsely (sb-MIL) composed of the bag. However, for fine-grained emotion recognition, we do not know whether the post-stimuli emotions are the most salient (only small amount of the instances are correlated with the post-stimuli label) or overall (most of the instances are correlated with the post-stimuli label) emotions of users. That makes it challenging for classic MIL methods to identify the instances which are correlated with the post-stimuli labels for fine-grained emotion recognition.

For the deep learning based weakly-supervised methods, we find that all three of them provide better classification (average acc +7.75%) and regression (average mse -0.097)) results compared with the four classic MIL methods. However, we also find out that all three methods result in problems of overfitting for the classification task. The accuracies on training sets are much higher than the accuracies on testing sets: the accuracy differences for training and testing sets are 23.14%, 19.27% and 22.28% for AMIL, WSPPG and WCRNN, respectively. The results demonstrate that deeper network or more complex structures (i.e., attention structure and recurrent structure) can decrease the generalizability of the algorithms by providing more accurate results only on the training set.

EDMIL obtains good recognition accuracy and w-f1 score ($>$ 65% accuracy and 0.65 w-f1) for all three datasets. By taking advantage of the end-to-end structure, *EDMIL* automatically obtains the matching scores for instances and bag labels without a pre-set hypothesis. Compared with classic MIL methods, we do not need to know whether most or only a small amount of the instances are correlated with the post-stimuli labels. Compared with the three baselines which use an end-to-end learning structure, *EDMIL* also achieves better performance (average acc +10.34%, mse -0.03). *EDMIL* does not suffer from overfitting: we only find the training accuracy is 3.46% (averaged from three datasets) higher than the testing accuracy for *EDMIL*. That is a result of the shallow feature extraction network and simple instance regularization module we use to design *EDMIL*. The result also shows that more accurate fine-grained emotion recognition can be achieved using deep neural network based (compared with traditional machine learning based) weakly-supervised learning algorithms.

5.5 Ablation Study

We conduct an ablation study to verify the effectiveness of each component. Since our algorithm needs MIL layers to obtain fine-grained V-A, we begin with only using the MIL layers to train the network. The MIL layers directly use the raw signal segments without passing them through the pre-

TABLE 4
Ablation Study for Pre-Processing (PP), Feature Extraction (FE) and Multiple Instance Learning (MIL) Module

		CASE			MERCA			CEAP-360VR		
		acc	w-f1	mse	acc	w-f1	mse	acc	w-f1	mse
MIL	valence	56.12%	0.53	0.4052	51.71%	0.49	0.3956	49.52%	0.41	0.4215
	arousal	51.57%	0.49	0.4132	50.12%	0.47	0.4056	53.61%	0.45	0.4372
PP+MIL	valence	58.21%	0.51	0.4011	53.38%	0.53	0.3327	51.26%	0.50	0.4112
	arousal	54.38%	0.49	0.3981	52.32%	0.51	0.3152	55.21%	0.51	0.4009
FE+MIL	valence	72.52%	0.71	0.2654	59.67%	0.57	0.2738	61.26%	0.59	0.2953
	arousal	75.66%	0.72	0.2477	57.21%	0.56	0.2235	63.14%	0.59	0.2817
PP+FE+MIL	valence	**75.93%**	**0.72**	**0.2354**	**70.51%**	**0.69**	**0.2673**	**65.04%**	**0.65**	**0.2384**
	arousal	**79.73%**	**0.77**	**0.2281**	**67.62%**	**0.65**	**0.2051**	**67.05%**	**0.65**	**0.2529**

processing and feature extraction module. Then we test the performance of combining the MIL layers with the pre-processing (PP) and feature extraction (FE) layers respectively. Finally, we combine all the modules in *EDMIL* and present the results for comparison.

As shown in Table 4, both FE and PP contribute to the classification and regression tasks. The FE benefits the network by extracting deep features for MIL layers to learn the probability for instances to predict the corresponding post-stimuli labels. Thus, the recognition accuracies increase 12.80% and mse drops 0.148 on average after combining FE to MIL. The increased performance of adding PP is not as significant as adding FE: the recognition accuracies increase 6.07% and mse drops 0.027 on average after combining PP to the network. The reason of this is that the convolution layers of FE have already automatically filtered some of the noise and artifacts in the signals when extracting the features. In conclusion, all components contribute to both the classification and regression tasks. The observations above demonstrate the effectiveness of the components in the proposed algorithm.

5.6 Comparison Between Different Feature Extraction Methods

As we introduced in section 3.2, we compare three feature extraction methods (*deepfeat, pcorrfeat and manualfeat*) in the feature extraction layer of *EDMIL*. The purpose of this comparison is to find out whether the deep features (*deepfeat*) learned by the end-to-end neural network can provide more accurate classification and regression results compared with unsupervised feature extraction method (*pcorrfeat*) and manually selected features (*manualfeat*).

As shown in Fig. 8, *deepfeat* results in the highest accuracy. *pcorrfeat* and *manualfeat* have similar accuracy which are lower than *deepfeat*. In addition, the accuracy on the training and testing set is similar (training accuracies are 2.6%, 3.7% and 4.1% higher than testing for CASE, MERCA and CEAP-360VR respectively). The results indicate that using an end-to-end feature extraction method does not result in overfitting (found by previous works on fully-supervised learning for emotion recognition [3], [69]) due to the weakly-supervised structure we use.

However, the mse of *pcorrfeat* and *manualfeat* are lower than *deepfeat*. Thus, using unsupervised and manually extracted features can result in better regression results. The reason of higher mse for *deepfeat* is that the deep features are learned for the classification task, not the regression task. The *pcorrfeat* and *manualfeat* however, are not constrained with the classification task, which can better represent the dynamic patterns of V-A changes. However, *EDMIL* is designed for the classification task particularly and the post-stimuli emotion labels for training are also the discretized labels. In addition, for the regression task, the maximum and minimum V-A are needed to normalize the instance gains. It means users are expected to input their highest and lowest V-A during the whole video watching, which is sometimes difficult for users if the video is long. Thus, the classification is more applicable (users only need to input their post-stimuli V-A) in real-life scenarios.

6 Discussion

6.1 Fully-Supervised Versus Weakly Supervised: Advantages and Disadvantages

The baseline methods we test in section 5.4 are all weakly supervised methods trained with post-stimuli emotion labels. The fully-supervised learning methods, however, learn the instance-label relationship by building the direct mapping between instances and fine-grained emotion labels. Thus, it is interesting to compare the results between training with post-stimuli labels (weakly-supervised) and fine-grained labels

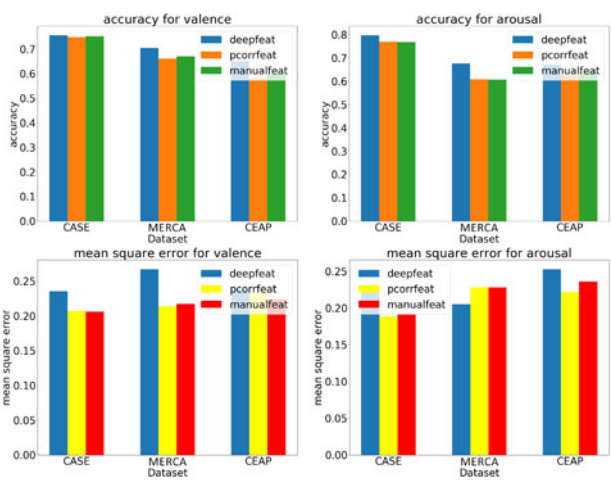

Fig. 8. The accuracy and mse of deepfeat, pcorrfeat and manualfeat on three datasets.

TABLE 5
The Comparison Between Weakly-Supervised (*EDMIL*) and Fully-Supervised (*1D-CNN And LSTM*) Methods

		EDMIL		1D-CNN [42]		LSTM [108]	
		acc	dtw	acc	dtw	acc	dtw
CASE	**valence**	**75.63%**	16.0417	53.04%	**10.886**	54.74%	11.77
	arousal	**79.73%**	12.6875	52.34%	**6.997**	53.82%	7.267
MERCA	**valence**	**70.51%**	11.7688	51.57%	6.183	53.88%	**6.031**
	arousal	**67.62%**	10.7917	42.91%	8.437	46.26%	**8.215**
CEAP-360VR	**valence**	**65.04%**	9.9973	47.33%	**5.941**	44.36%	6.021
	arousal	**67.05%**	9.3810	45.67%	**5.733**	43.27%	5.938

(fully-supervised). The comparison can help us understand whether the additional fine-grained (i.e., instance-level) labels can improve or compromise the performance of fine-grained emotion recognition.

To implement this comparison, we choose two widely used deep learning models, 1D-Convolutional Neural Network (*1D-CNN*) and Long Short Term Memory networks (LSTM) for comparison. We choose the basic *1D-CNN* [42] and *LSTM* networks [108] from previous works for emotion recognition to avoid over-tuning. We use the mean V-A label for each instance and directly train the *1D-CNN* and *LSTM* at instance-level (i.e., each instance has a corresponding V-A label). We run the 3-class classification task as we did for testing *EDMIL*. To evaluate performance, we use two metrics: the recognition accuracy and dynamic time warping distance (DTW) [109]. DTW is one of the most prominent methods in similarity measurement for time series data [110]. The results for the comparison are shown in Table 5.

As shown in Table 5, the fully-supervised algorithms achieve lower recognition accuracy compared with our weakly-supervised algorithm (*EDMIL*). The results of fully-supervised methods are supposed to be better than *EDMIL* since the fully-supervised algorithms have additional information (i.e., the instance-level labels) for training. We then compare the training and testing accuracy for both fully-supervised algorithms and *EDMIL*. We find that both the *1D-CNN* and *LSTM* have the problem of overfitting. The accuracies on training sets are much higher than the accuracies on testing sets (average of the three datasets: 20.04% and 18.28% higher for *1D-CNN* and *LSTM* respectively). However, for weakly-supervised training, we do not find a much higher accuracy on the training set (average of the three datasets: 3.46% higher for *EDMIL*). The overfitting can be a result of the temporal resolution mismatch between physiological signals and fine-grained self-reports. When annotating their emotions in real-time, different users have different awareness (interoception) levels about their emotions [70]. The relationships between instances and labels are different among users because the interoception levels across individuals are different. Thus, the recognition algorithm can learn contradictory information if we directly build a strong constraint between the fine-grained labels and signals. That also explains the finding of Romeo *et al.* [19] and Kandemir *et al.* [67] that building a subject-independent emotion recognition model is challenging, especially for fine-grained emotion recognition.

Although recognition accuracies are lower, the DTW distances of the fully-supervised methods are lower than *EDMIL*. Lower DTW distance means higher similarity of two time sequences. That indicates that the fully-supervised algorithms can result in better recognition results for the whole video instead of individual instances. Compared with accuracy, DTW is less sensitive to the time-shift of specific values in the sequence. Fig. 9 shows three examples of the prediction results of *1D-CNN* and *EDMIL* on three datasets. Taking the example of Fig. 9c, although *EDMIL* achieves higher accuracy (86.21% v.s. 55.17%) in this specific case, the DTW of *EDMIL* is higher (6.0 v.s. 1.0) than *1D-CNN*. The results also show that there is temporal mismatch between individuals: when the evaluation metric (i.e., DTW) is less sensitive to the time-shift, fully-supervised methods have better performance. Since we run subject-independent validation, the temporal relationship between input signals and emotions learned from other subjects is different from the one used for testing. That causes shifts of predictions in the time domain, which makes the instance-level accuracies low but does not effect the sequence-level prediction.

We also add the original signals to Fig. 9. From Fig. 9 we can see that the arousal labels have a clear correlation between the EDA (Figs. 9a and 9c), which is in line with most of the studies of previous works [90]. The heart rate and skin temperature correlate with both the valence and arousal changes [111], [112]. We also find that some of the changes which are ignored by *EDMIL* but captured by the fully-supervised algorithm are also shown in the physiological signals. For example, for Fig. 9a, there is a duration of high arousal predicted by the fully-supervised algorithm, which correlates to an increase in heart rate. However, *EDMIL* predicts the emotion to be neutral during this duration. The visual comparison between signals and predictions validate our conclusion that fully-supervised algorithms can result in better recognition results for whole videos instead of individual instances.

6.2 Towards Temporally Precise Emotion Recognition: How Fine-Grained Can the Recognition Be?

The performance of *EDMIL* is influenced by the structure of the bag: the length of the instance can affect the accuracy and the temporal resolution of recognition [3]. The shorter instance length representing a higher number of instances for one video watching will lead to a finer level of

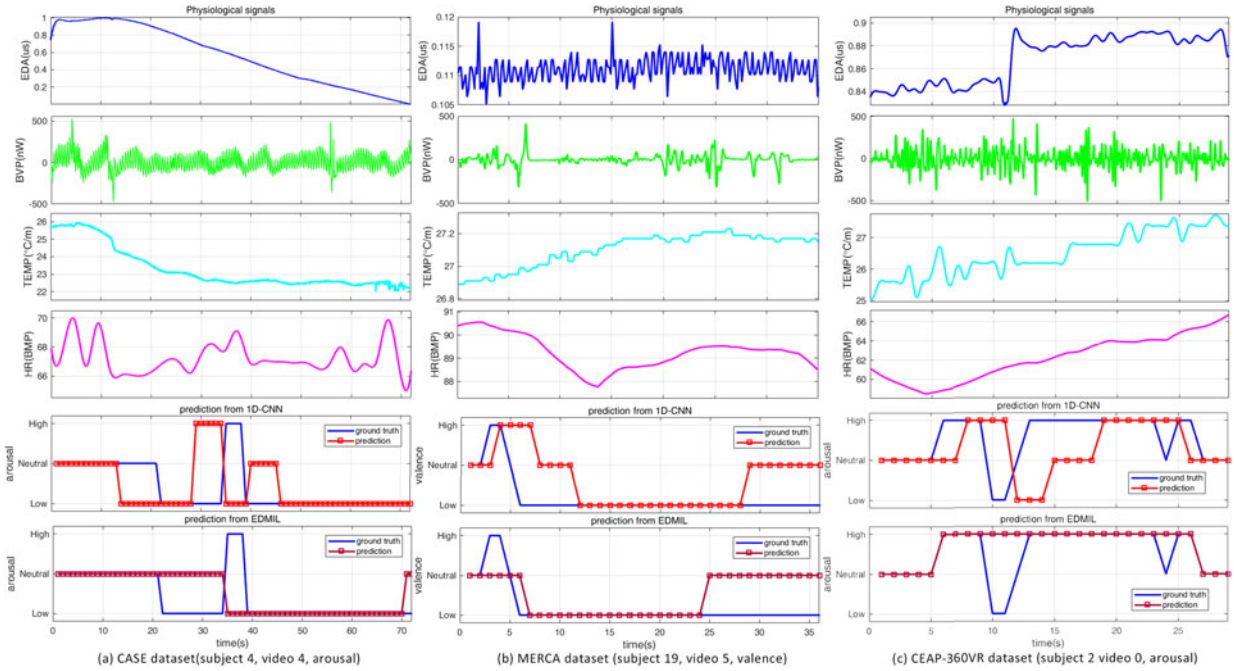

Fig. 9. Examples of physiological signals and prediction results for fully-supervised (*1D-CNN*) and weakly-supervised (*EDMIL*) algorithm.

granularity in temporal resolution. However, a too-short instance length can bring challenges for feature extraction because the information inside each instance can be insufficient for accurate recognition [3], [19]. Thus, it is worthwhile to find out what are the appropriate instance lengths for fine-grained emotion recognition based on deep multiple instance learning. Thus, we conduct an experiment to test *EDMIL* with different instance lengths on CASE, MERCA, and CEAP-360VR, respectively. The recognition accuracies of different instance lengths are shown in Fig. 10.

As shown in Fig. 10, the recognition accuracy decreases significantly if the instance length is $\geq 5s$. This result is coherent with the finding from Romeo *et al.* [19] that a low number of longer instances may lose salient information related to the local physiological response. The accuracy also decreases when the instance length is $< 1s$. Since emotion states are classified based on the features from each instance, a short instance length can entail insufficient information for accurate classification [3]. The instance length of 2s achieves the best recognition accuracy for both valence and arousal. For all the datasets, instance length from $1s$ to $2s$ can result in good recognition results (up to 60%). The result is in line with the research from Paul *et al.* [6] that the duration of emotion typically ranges from 0.5s to 4s. The takeaway message from this experiment is that an instance length between $1s$ to $2s$ is the appropriate length for fine-grained emotion recognition using physiological signals and deep multiple instance learning.

6.3 Does the Percentage of Post-Stimuli Annotations Affect Recognition Accuracy?

The performance of *EDMIL* can also be influenced by the percentage of post-stimuli V-A users annotated in each session. Traditional MIL methods require pre-set hypotheses about whether the instances corresponding to the post-stimuli annotation densely or sparsely consist of the bag [59], [62]. For example, if a user annotates that he or she experienced happiness after watching one video, traditional MIL methods can only obtain accurate predictions if the user felt happy most of the time when watching the video. In addition, if all (or most) of the instances are annotated the same as the post-stimuli annotation, we can just do bag-level prediction and train all the instances using fully-supervised learning. In that case, it is not needed to develop weakly-supervised learning algorithms for identifying the instances which contribute to predicting the post-stimuli annotation.

To circumvent this, we first check the percentage of instances which are annotated the same as the post-stimuli annotations for the three datasets we use. For each dataset, suppose there are S users who watch L videos. For each session, the user annotates one post-stimuli V-A. Meanwhile, the user also annotates the fine-grained V-A for K instances inside this session. The number of instances which the user annotated the same as the post-stimuli annotations are N. The percentage of the post-stimuli annotation for this session is $p = N/K$. Then for the $S \times L$ sessions we obtain $P = [p_1, p_2, \ldots, p_{S \times L}]$ of V-A

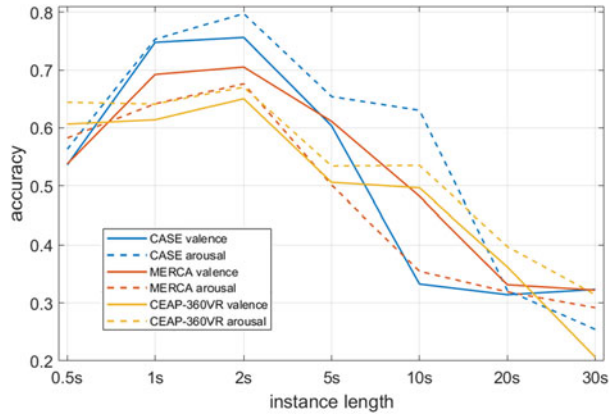

Fig. 10. The recognition accuracy by different instance lengths on CASE, MERCA and CEAP-360VR.

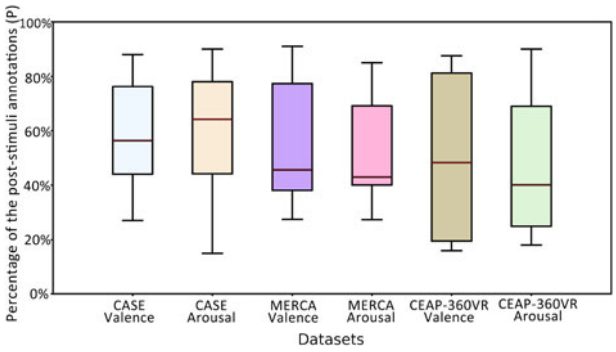

Fig. 11. The percentage of the post-stimuli annotations in each session (one user watches one video) for CASE, MERCA and CEAP-360VR.

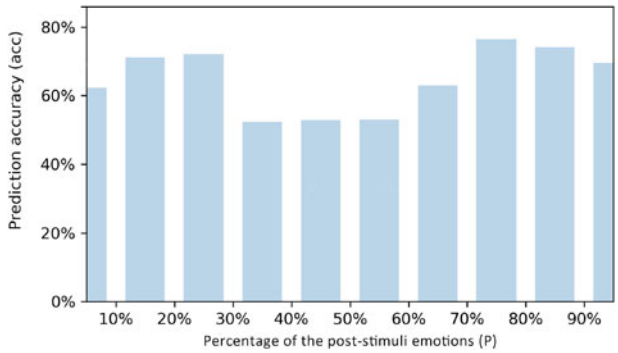

Fig. 12. The relationship between the percentage of the post-stimuli annotations and recognition accuracy.

for the whole dataset. As shown in Fig. 11, the mean and standard deviation of P for three datasets are: CASE-valence: $58.0\%(0.18)$, MERCA-valence: $53.5\%(0.20)$, CEAP-360VR-valence: $49.5\%(0.27)$, CASE-arousal: $57.4\%(0.18)$, MERCA-arousal: $50.2\%(0.18)$, CEAP-360VR-arousal: $46.5\%(0.25)$. A Shapiro-Wilk tests shows that the percentages of the post-stimuli annotations for all three datasets are not normally distributed (all $p < 0.05$ for three datasets). As we compare three unmatched groups, we perform a Kruskal-Wallis [113] test. We find significant differences of the percentages of the post-stimuli valence ($\chi^2(2) = 10.97, p < 0.05$) and arousal ($\chi^2(2) = 35.29, p < 0.05$) among the three datasets. We then run post-hoc pairwise comparison tests using Mann-Whitney test [114] with Bonferroni correction. The p-values and effect sizes for the pairwise comparison are shown in Table 6. In the tests, we find pairwise significant differences (all $p < 0.05$) of P for both valence and arousal between datasets. The results demonstrate that the datasets contain sessions with different levels of time ambiguity (the percentage of the post-stimuli annotations for different datasets are significantly different), which makes it challenging for recognition algorithms to have a generalizable performance on all three datasets.

To find out whether the percentage of post-stimuli annotation influences the recognition accuracy, we calculate the average accuracy of V-A and the percentage of the post-stimuli annotations for all the sessions in the three datasets. As shown in Fig. 12, *EDMIL* achieves up to 60% of the recognition accuracy if the post-stimuli annotation accounts for more than 60% or less than 30% of one session. If the post-stimuli annotation accounts for 30% to 60% of one session, the accuracy is lower but still more than 55%. Although the result shows up to 50% of accuracy for all the sessions, the performance of *EDMIL* still decreases by around 10% when the post-stimuli annotation is neither densely (more than 60%) nor sparsely (less than 30%) consists of the bag. The deep structure of *EDMIL* can automatically determine whether the post-stimuli annotation densely or sparsely consists of the bag. Thus, for these two conditions, *EDMIL* can achieve relatively high accuracy. However, *EDMIL* recognizes the post-stimuli annotation for each instance based on the probability that the instance matches the post-stimuli annotation. Similar to traditional MIL methods, when the post-stimuli annotation neither densely nor sparsely consists of the bag, the probabilities for instances tend to be similar [115], which makes it difficult for the network to identify the corresponding instances. The takeaway message of this experiment is that the percentage of the post-stimuli annotations do have an influence on the recognition accuracy. *EDMIL* can achieve the highest recognition accuracy if users annotate their most salient but short emotions (less than 30%), or their overall and longer (i.e., persisting) emotions (more than 60%) after watching the video.

TABLE 6
The Post-Hoc Pairwise Comparison Tests Using Mann-Whitney Test on the Percentages of the Post-Stimuli Valence and Arousal

(a) valence

p-value \ effect size	CASE	MERCA	CEAP-360VR
CASE		0.566	0.583
MERCA	0.026		0.553
CEAP-360VR	0.004	0.038	

(b) arousal

p-value \ effect size	CASE	MERCA	CEAP-360VR
CASE		0.611	0.650
MERCA	<0.001		0.576
CEAP-360VR	<0.001	0.003	

7 Limitations and Future Work

Given the challenges of predicting valence and arousal labels at a fine level of granularity using only post-stimuli labels, there are naturally limitations to our work. First, *EDMIL* can only identify the annotated (post-stimuli) emotion from the baseline emotion (e.g., neutral) because only post-stimuli labels are used for training. The none-annotated emotions are all categorized as part of the baseline because of their low matching score for predicting the post-stimuli labels. In addition, the regression performance of *EDMIL* is not as good as classification since the network is designed specifically for classification. In the future, we will extend the algorithm into a multi-instance multi-label formulation [116]. Secondly, we do not test *EDMIL* on longer duration stimuli because of the limited availably of datasets. The video lengths of emotion recognition datasets are commonly short (usually $<$ 3 mins [15], [22], [23]) to avoid (visual) fatigue of participants. We will test the performance of *EDMIL* for longer stimuli in the future when more

datasets are available. Moreover, the users in our study were all adults. Previous works [117], [118] have shown that users of different ages may have different emotional reactions to the same video. We did not test *EDMIL* on elderly or children since there are limited amount of datasets which contain continuously annotated physiological signals from users across age groups. In the future, we plan to test our algorithm on users of different ages if more datasets become publicly available.

Another limitation of our work is that we only consider physiological signals as the input modality. The application scenario of our paper is to analyze the personalized experience (emotions) of users while watching videos. Thus, the semantic features (e.g., audio-visual content, text caption of the content, speech transcripts) from video stimuli can also help to improve the recognition results by providing the context information for recognition. The text-derived fingerprints are more accessible to generate this context information compared with video data because of its low cost of computational recourses [119]. In the future, we can build a weak constraint (e.g., using Canonical Correlation Analysis (CCA) [3]) between the emotion semantic features derived from the text of videos (e.g., speech transcripts, emotion tags of videos) to promote the recognition accuracy.

At last, our algorithm can help video providers to build an emotion analytics dashboard by only asking users to annotate their post-stimuli emotions after one video watching. By adding an emotion layer to the videos, our algorithm can help video providers to understand the dynamic changes of users' emotions toward their products and adjust the content based on that. The predicted emotions will be aligned with the video content and visualized on the dashboard for video providers to analyze the relationship between video content and the emotions of users. In the future, we plan to apply our algorithm to such emotion analytics dashboard for an application in real world scenario.

8 Conclusion

Fine-grained emotion recognition requires training the algorithm with fine-grained emotion ground truth labels to build the mapping between segments of signals and corresponding emotions. In this paper, we propose *EDMIL*, a deep multiple instance learning based emotion recognition algorithm to classify fine-grained valence and arousal trained with only post-stimuli emotion labels. The algorithm uses weakly-supervised learning to model the temporal ambiguity of post-stimuli emotion labels and learn the instance-label relationship according to the probability for each instance to predict the post-stimuli label. The proposed algorithm achieves reasonable performance (more than 65% on high/neutral/low classification) for subject-independent testing on three datasets collected in three different environments (i.e., desktop, mobile, and HMD-based VR). *EDMIL* also outperforms the classic multiple instance learning methods which previous work [19] used for emotion recognition. Running tests on three different datasets, we found that *EDMIL* achieves similar recognition accuracy in desktop, mobile and VR environments, which indicates its good generalizable performance. Finally, our experiments show that (1) weakly supervised learning can reduce overfitting caused by the temporal mismatch between fine-grained annotations and input signals, (2) instance segment lengths between 1-2s result in the highest recognition accuracies, (3) *EDMIL* can achieve the highest recognition accuracy if users annotate their most salient but short emotions, or their overall and longer-duration (i.e., persisting) emotions, and (4) feature extraction using an end-to-end structure can improve recognition accuracy compared with manual feature extraction as well as unsupervised feature extraction methods.

References

[1] S. Khorram, M. McInnis, and E. M. Provost, "Jointly aligning and predicting continuous emotion annotations," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1069–1083, Oct.–Dec. 2021.

[2] M. Abdul-Mageed and L. Ungar, "Emonet: Fine-grained emotion detection with gated recurrent neural networks," in *Proc. 55th Annu. meeting Assoc. Comput. Linguistics*, 2017, pp. 718–728.

[3] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, "CorrNet: Fine-grained emotion recognition for video watching using wearable physiological sensors," *Sensors*, vol. 21, no. 1, 2021, Art. no. 52.

[4] F. Hasanzadeh, M. Annabestani, and S. Moghimi, "Continuous emotion recognition during music listening using EEG signals: A fuzzy parallel cascades model," 2019, *arXiv:1910.10489*.

[5] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 90–100, Mar. 2006.

[6] E. Paul, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Toronto, Canada: OWL Books, 2007.

[7] R. W. Levenson, "Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity," *Social Psychophysiology: Theory Clin. Appl.*, New York, NY, USA: Wiley, 1988.

[8] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmüller, "Emujoy: Software for continuous measurement of perceived emotions in music," *Behav. Res. Methods*, vol. 39, no. 2, pp. 283–290, 2007.

[9] T. Zhang, A. El Ali , C. Wang, A. Hanjalic, and P. Cesar, "RCEA: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–15.

[10] M. Soleymani, S. Asghari-Esfeden , Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, First Quarter 2015.

[11] S. Wu, Z. Du, W. Li, D. Huang, and Y. Wang, "Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze," in *Proc. Int. Conf. Multimodal Interaction*, 2019, pp. 40–48.

[12] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 176–183.

[13] G. Van Houdt , C. Mosquera, and G. Napoles, "A review on the long short-term memory model," *Artif. Intell. Rev.*, vol. 53, pp. 5929–5955, 2020.

[14] J. J. Rivas *et al.*, "Unobtrusive inference of affective states in virtual rehabilitation from upper limb motions: A feasibility study," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 470–481, Third Quarter 2020.

[15] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Sci. Data*, vol. 6, no. 1, pp. 1–13, 2019.

[16] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

[17] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2015: The 5th international audio/visual emotion challenge and workshop," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1335–1336.

[18] M. Valstar *et al.*, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 3–10.

[19] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil, "Multiple instance learning for emotion recognition using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 389–407, Oct.–Dec. 2022.

[20] M. Song *et al.*, "Audiovisual analysis for recognising frustration during game-play: Introducing the multimodal game frustration database," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interaction*, 2019, pp. 517–523.

[21] I. Abdic, L. Fridman, D. McDuff, E. Marchi, B. Reimer, and B. Schuller, "Driver frustration detection from audio and video in the wild," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1354–1360.

[22] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012.

[23] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 42–55, Jan.–Mar. 2012.

[24] Y. Liu and O. Sourina, "Eeg databases for emotion recognition," in *Proc. Int. Conf. Cyberworlds*, 2013, pp. 302–309.

[25] B. L. Fredrickson and D. Kahneman, "Duration neglect in retrospective evaluations of affective episodes." *J. Pers. social Psychol.*, vol. 65, no. 1, 1993, Art. no. 45.

[26] J. Wu, Z. Zhou, Y. Wang, Y. Li, X. Xu, and Y. Uchida, "Multi-feature and multi-instance learning with anti-overfitting strategy for engagement intensity prediction," in *Proc. Int. Conf. Multimodal Interaction*, 2019, pp. 582–588.

[27] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. London, U.K.: Chapman and Hall/CRC, 2012.

[28] J. A. Russell, "A circumplex model of affect." *J. Pers. social Psychol.*, vol. 39, no. 6, 1980, Art. no. 1161.

[29] L. Shu *et al.*, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, 2018, Art. no. 2074.

[30] T. Xue, A. El Ali, T. Zhang, G. Ding, and P. Cesar, "RCEA-360VR: Real-time, continuous emotion annotation in 360VR videos for collecting precise viewport-dependent ground truth labels," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–15.

[31] W. Yang, M. Rifqi, C. Marsala, and A. Pinna, "Towards better understanding of player's game experience," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2018, pp. 442–449.

[32] S. Wioleta, "Using physiological signals for emotion recognition," in *Proc. 6th Int. Conf. Hum. System Interact.*, 2013, pp. 556–561.

[33] X. Niu, L. Chen, H. Xie, Q. Chen, and H. Li, "Emotion pattern recognition using physiological signals," *Sensors Transducers*, vol. 172, no. 6, 2014, Art. no. 147.

[34] E. Di Lascio, S. Gashi, and S. Santini, "Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–21, 2018.

[35] M. Zecca, S. Micera, M. C. Carrozza, and P. Dario, "Control of multifunctional prosthetic hands by processing the electromyographic signal," *Crit. Reviews™ Biomed. Eng.*, vol. 30, no. 4–6, pp, 459–485, 2002.

[36] S. Huynh, S. Kim, J. Ko, R. K. Balan, and Y. Lee, "Engagemon: Multi-modal engagement sensing for mobile games," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–27, 2018.

[37] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, "Personalized emotion recognition by personality-aware high-order learning of physiological signals," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 1s, pp. 1–18, 2019.

[38] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affective Comput.*, vol. 9, no. 2, pp. 147–160, Apr.–Jun. 2016.

[39] A. Raheel, M. Majid, and S. M. Anwar, "DEAR-MULSEMEDIA: Dataset for emotion analysis and recognition in response to multiple sensorial media," *Inf. Fusion*, vol. 65, pp. 37–49, 2021.

[40] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological Signals Based Human Emotion Recognition: A Review," in *Proc. IEEE 7th Int. Colloquium Signal Process. Appl.*, 2011, pp. 410–415.

[41] M. Ali, F. Al Machot, A. H. Mosa, and K. Kyamakya, "CNN based subject-independent driver emotion recognition system involving physiological signals for ADAs," in *Proc. Adv. Microsyst. Automo. Appl.*, 2016, pp. 125–138.

[42] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos)," *IEEE Access*, vol. 7, pp. 57–67, 2018.

[43] S.-h. Zhong, A. Fares, and J. Jiang, "An attentional-LSTM for improved classification of brain activities evoked by images," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1295–1303.

[44] X. Zhang *et al.*, "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4386–4399, Sep. 2021.

[45] T. Zhang, "Multi-modal fusion methods for robust emotion recognition using body-worn physiological sensors in mobile environments," in *Proc. Int. Conf. Multimodal Interaction*, 2019, pp. 463–467.

[46] C.-Y. Chang, J.-Y. Zheng, and C.-J. Wang, "Based on support vector regression for emotion recognition using physiological signals," in *Proc. Int. Joint Conf. Neural Netw.*, 2010, pp. 1–7.

[47] J. Wei, T. Chen, G. Liu, and J. Yang, "Higher-order multivariable polynomial regression to estimate human affective states," *Sci. Rep.*, vol. 6, 2016, Art. no. 23384.

[48] M. Awais *et al.*, "LSTM based emotion detection using physiological signals: Iot framework for healthcare and distance learning in COVID-19," *IEEE Internet Things J.*, vol. 8, no. 23, pp. 16863–16871, Dec. 2021.

[49] A. Srinivasan, S. Abirami, N. Divya, R. Akshya, and B. Sreeja, "Intelligent child safety system using machine learning in iot devices," in *Proc. 5th Int. Conf. Comput., Commun. Secur.*, 2020, pp. 1–6.

[50] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affective Comput.*, vol. 2, no. 2, pp. 92–105, Apr.–Jun. 2011.

[51] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *Proc. ISCA Tutorial Res. Workshop Speech Emotion*, 2000.

[52] K. Sharma, C. Castellini, F. Stulp, and E. L. Van den broek, "continuous, real-time emotion annotation: A novel joystick-based analysis framework," *IEEE Trans. Affective Comput.*, vol. 11, no. 1, pp. 78–84, Jan.–Mar. 2020.

[53] D. Lottridge and M. Chignell, "Sliders rate valence but not arousal: Psychometrics of self-reported emotion assessment," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 54. Newbury Park, CA, USA: SAGE, 2010, pp. 1766–1770.

[54] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognit.*, vol. 77, pp. 329–353, 2018.

[55] J. Feng and Z.-H. Zhou, "Deep miml network," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1884–1890.

[56] T. Rao, M. Xu, H. Liu, J. Wang, and I. Burnett, "Multi-scale blocks based image emotion classification using multiple instance learning," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 634–638.

[57] N. Pappas and A. Popescu-Belis, "Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 455–466.

[58] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, P. G. Georgiou, and S. S. Narayanan, "Affective state recognition in married couples' interactions using pca-based vocal entrainment measures with multiple instance learning," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2011, pp. 31–41.

[59] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *Proc. Int. Conf. Mach. Learn.*, 2002, Art. no. 7.

[60] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Proc. Adv. neural Inf. Process. Syst.*, 2006, pp. 1417–1424.

[61] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 1073–1080.

[62] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 105–112.

[63] X. Zhang *et al.*, "Emotion recognition based on electroencephalogram using a multiple instance learning framework," in *Proc. Int. Conf. Intell. Comput.*, 2018, pp. 570–578.

[64] Y. Tian, W. Hao, D. Jin, G. Chen, and A. Zou, "A review of latest multi-instance learning," in *Proc. 4th Int. Conf. Comput. Sci. Artif. Intell.*, 2020, pp. 41–45.

[65] Y. Wang *et al.*, "Automatic depression detection via facial expressions using multiple instance learning," in *Proc. IEEE 17th Int. Symp. Biomed. Imag.*, 2020, pp. 1933–1936.

[66] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 47–56.

[67] M. Kandemir, A. Vetek, M. Goenen, A. Klami, and S. Kaski, "Multi-task and multi-view learning of user state," *Neurocomputing*, vol. 139, pp. 97–106, 2014.

[68] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[69] T. Zhang, A. El Ali, C. Wang, X. Zhu, and P. Cesar, "Corrfeat: Correlation-based feature extraction algorithm using skin conductance and pupil diameter for emotion recognition," in *Proc. Int. Conf. Multimodal Interaction*, 2019, pp. 404–408.

[70] H. D. Critchley and S. N. Garfinkel, "Interoception and emotion," *Curr. Opin. Psychol.*, vol. 17, pp. 7–14, 2017.

[71] D. Kukolja, S. Popović, M. Horvat, B. Kovač, and K. Ćosić, "Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications," *Int. J. Hum.- Comput. Stud.*, vol. 72, no. 10/11, pp. 717–727, 2014.

[72] A. v. d. Oord *et al.*, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[73] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters-improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4353–4361.

[74] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[75] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[76] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[77] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biol. Psychol.*, vol. 84, no. 3, pp. 394–421, 2010.

[78] J. R. Loaiza, "Emotions and the problem of variability," *Rev. Philosophy Psychol.*, vol. 12, pp. 329–351, 2021.

[79] H. J. Nussbaumer, "The Fast Fourier Transform," in *Fast Fourier Transform and Convolution Algorithms*. Berlin, Germany: Springer, 1981, pp. 80–111.

[80] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 3376–3391.

[81] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu, "A sufficient condition for convergences of adam and RMSProp," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11127–11135.

[82] T. Xue, A. El Ali, T. Zhang, G. Ding, and P. Cesar, "CEAP-360VR: A continuous physiological and behavioral emotion annotation dataset for 360 videos," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2021.3124080.

[83] T. T. Lin and C. Chiu, "Investigating adopter categories and determinants affecting the adoption of mobile television in china," *China Media Res.*, vol. 10, no. 3, pp. 74–87, 2014.

[84] J. McNally and B. Harrington, "How millennials and teens consume mobile video," in *Proc. ACM Int. Conf. Interactive Experiences TV Online Video*, 2017, pp. 31–39.

[85] K. O'Hara, A. S. Mitchell, and A. Vorbau, "Consuming video on mobile devices," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2007, pp. 857–866.

[86] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[87] B. J. Li, J. N. Bailenson, A. Pines, W. J. Greenleaf, and L. M. Williams, "A public database of immersive VR videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures," *Front. Psychol.*, vol. 8, 2017, Art. no. 2116.

[88] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[89] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *J. Neural Eng.*, vol. 16, no. 3, 2019, Art. no. 031001.

[90] J. Shukla, M. Barreda-Angeles, J. Oliver, G. Nandi, and D. Puig, "Feature extraction and selection for emotion recognition from electrodermal activity," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 857–869, Oct.–Dec. 2021.

[91] K. Gouizi, F. Bereksi Reguig, and C. Maaoui, "Emotion recognition from physiological signals," *J. Med. Eng. Technol.*, vol. 35, no. 6–7, pp. 300–307, 2011.

[92] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, "Recognizing emotions induced by affective sounds through heart rate variability," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 385–394, Oct.–Dec. 2015.

[93] J. Fleureau, P. Guillotel, and I. Orlac, "Affective benchmarking of movies based on the physiological responses of a real audience," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 73–78.

[94] Y. Chu, X. Zhao, J. Han, and Y. Su, "Physiological signal-based method for measurement of pain intensity," *Front. Neurosci.*, vol. 11, 2017, Art. no. 279.

[95] P. Karthikeyan, M. Murugappan, and S. Yaacob, "Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress," *J. Phys. Ther. Sci.*, vol. 24, no. 12, pp. 1341–1344, 2012.

[96] E. Meijering, "A chronology of interpolation: From ancient astronomy to modern signal and image processing," *Proc. IEEE*, vol. 90, no. 3, pp. 319–342, Mar. 2002.

[97] R. W. Daniels, *Approximation Methods for Electronic Filter Design: With Applications to Passive, Active, and Digital Networks*. New York, NY, USA: McGraw-Hill, 1974.

[98] P. Van Gent, H. Farah, N. Nes, and B. van Arem, "Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data," in *Proc. 6th HUMANIST Conf.*, 2018, pp. 173–178.

[99] A. T. Zhang and B. O. Le Meur, "How old do you look? Inferring your age from your gaze," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 2660–2664.

[100] L. Prechelt, "Early stopping-but when?," in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 1998, pp. 55–69.

[101] H. Ferdinando and E. Alasaarela, "Enhancement of emotion recogniton using feature fusion and the neighborhood components analysis." in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2018, pp. 463–469.

[102] N. Chinchor, "Muc-3 evaluation metrics," in *Proc. 3rd Conf. Message Understanding*, 1991, pp. 17–24.

[103] M. Fatourechi, R. K. Ward, S. G. Mason, J. Huggins, A. Schlögl, and G. E. Birch, "Comparison of evaluation metrics in classification applications with imbalanced datasets," in *Proc. 7th Int. Conf. Mach. Learn. Appl.*, 2008, pp. 777–782.

[104] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 577–584.

[105] S. Mao, P. Ching, C.-C. J. Kuo, and T. Lee, "Advancing multiple instance learning with attention modeling for categorical speech emotion recognition," 2020, *arXiv:2008.06667*.

[106] J. Du, S.-Q. Liu, B. Zhang, and P. C. Yuen, "Weakly supervised rppg estimation for respiratory rate estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2391–2397.

[107] Y. Chen, H. Dinkel, M. Wu, and K. Yu, "Voice activity detection in the wild via weakly supervised sound event detection." in *Proc. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3665–3669.

[108] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 985–990.

[109] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, Feb. 1978.

[110] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and mining of time series data: Experimental comparison of representations and distance measures," *Proc. VLDB Endowment*, vol. 1, no. 2, pp. 1542–1552, 2008.

[111] H. Hägglund *et al.*, "Cardiovascular autonomic nervous system function and aerobic capacity in type 1 diabetes," *Front. Physiol.*, vol. 3, 2012, Art. no. 356.

[112] J. L. Greaney, W. L. Kenney, and L. M. Alexander, "Sympathetic regulation during thermal stress in human aging and disease," *Autonomic Neurosci.*, vol. 196, pp. 81–90, 2016.
[113] P. E. McKight and J. Najab, "Kruskal-wallis test," *The Corsini Encyclopedia Psychol.*, Hoboken, NJ, USA: Wiley, 2010.
[114] P. E. McKnight and J. Najab, "Mann-whitney U test," *The Corsini Encyclopedia Psychol.*, Hoboken, NJ, USA: Wiley, 2010.
[115] W. Zhang, L. Liu, and J. Li, "Robust multi-instance learning with stable instances," in *Proc. 24th Eur. Conf. Artif. Intell.*, 2020, pp. 718–725.
[116] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artif. Intell.*, vol. 176, no. 1, pp. 2291–2320, 2012.
[117] D. Y. Yeung, D. M. Isaacowitz, W. W. Lam, J. Ye, and C. L. Leung, "Age differences in visual attention and responses to intergenerational and non-intergenerational workplace conflicts," *Front. Psychol.*, vol. 12, 2021, Art. no. 2090.
[118] L. Fernández-Aguilar, J. Ricarte, L. Ros, and J. M. Latorre, "Emotional differences in young and older adults: Films as mood induction procedure," *Front. Psychol.*, vol. 9, 2018, Art. no. 1110.
[119] Z. Erenel, O. R. Adegboye, and H. Kusetogullari, "A new feature selection scheme for emotion recognition from text," *Appl. Sci.*, vol. 10, no. 15, 2020, Art. no. 5351.

Tianyi Zhang (Student Member, IEEE) is currently working toward the PhD degree with the faculty of Electrical Engineering, Mathematics & Computer Science (EEMCS) in Delft University of Technology. He is associated with the Distributed & Interactive Systems (DIS) group, Centrum Wiskunde & Informatica (CWI), the national research institute for mathematics and computer science in the Netherlands. His research interests lie in human-computer interaction and machine learning based affective computing.

Abdallah El Ali (Member, IEEE) received the PhD degree from the University of Amsterdam in 2013. He is currently a (tenured) research scientist with Centrum Wiskunde & Informatica (CWI) in Amsterdam within the Distributed & Interactive Systems (DIS) group. He is leading human-computer interaction (HCI) research with a focus on Affective Interactive Systems. His focus is on ground truth label acquisition techniques, affective state visualization across environments (mobile, wearable, XR), and bio-responsive interactive prototypes. For more information, please visit: https://abdoelali.com

Chen Wang received the PhD degree from the Vrije Universiteit Amsterdam, The Netherlands in 2018. She is currently a senior researcher and vice director with the Xinhuanet & State Key Laboratory of Media Convergence Production Technology and Systems, Future Media and Convergence Institute in Beijing, China. Her research interests include sensors technology, human computing interaction and graphical user interfaces.

Alan Hanjalic (Fellow, IEEE) is currently a professor of computer science, head with the Multimedia Computing Group and head with the Intelligent Systems Department, the Delft University of Technology (TU Delft), The Netherlands. His research interests are in the fields of multimedia information retrieval and recommender systems, in which he coauthored more than 250 publications. He is co-recipient of the Best Paper Award for the ACM Conference on Recommender Systems (ACM RecSys) 2012, the ACM International Conference on Multimedia (ACM Multimedia) 2017 and the IEEE International Conference on Multimedia Big Data (IEEE BigMM) 2019. He served as the chair of the Steering Committee of *IEEE Transactions on Multimedia*, the associate editor-in-chief of *IEEE MultiMedia Magazine*, and an associate editor of many scientific journals, including *IEEE Transactions in Multimedia*, *IEEE Transactions on Affective Computing* and *ACM Transactions on Multimedia Computing, Communications and Applications*. He also served as the general and/or program (co-)chair with the organizing committees of all major conferences in the multimedia domain, including ACM Multimedia, ACM CIVR/ICMR, and IEEE ICME.

Pablo Cesar (Senior Member, IEEE) leads the Distributed and Interactive Systems Group, Centrum Wiskunde & Infomartica (CWI) and is currently a professor with TU Delft, The Netherlands. His research combines human-computer interaction and multimedia systems, and focuses on modelling and controlling complex collections of media objects (including real-time media and sensor data) that are distributed in time and space. He has recently received the prestigious 2020 Netherlands Prize for ICT Research because of his work on human-centered multimedia systems. He is also the principal investigator from CWI on a number of projects on social virtual reality and affective computing. He is a member of the Editorial Board of *IEEE Multimedia*, *ACM Transactions on Multimedia*, and *IEEE Transactions of Multimedia*, among others. He has acted as an invited expert with the European Commission's Future Media Internet Architecture Think Tank. graphical user interfaces.