# Few-shot Learning for Fine-grained Emotion Recognition using Physiological Signals

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

**DOI**
10.1109/TMM.2022.3165715

**Publication date**
2022

**Document Version**
Final published version

**Published in**
IEEE Transactions on Multimedia

**Citation (APA)**
Zhang, T., El Ali, A., Hanjalic, A., & Cesar, P. (2022). Few-shot Learning for Fine-grained Emotion Recognition using Physiological Signals. *IEEE Transactions on Multimedia*, *25*, 3773-3787. Article 9751421. https://doi.org/10.1109/TMM.2022.3165715

# Few-Shot Learning for Fine-Grained Emotion Recognition Using Physiological Signals

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

***Abstract*—Fine-grained emotion recognition can model the temporal dynamics of emotions, which is more precise than predicting one emotion retrospectively for an activity (e.g., video clip watching). Previous works require large amounts of continuously annotated data to train an accurate recognition model, however experiments to collect such large amounts of continuously annotated physiological signals are costly and time-consuming. To overcome this challenge, we propose an Emotion recognition algorithm based on Deep Siamese Networks (EmoDSN) which can rapidly converge on a small amount of training data, typically less than 10 samples per class (i.e., <10 shot). EmoDSN recognizes fine-grained valence and arousal (V-A) labels by maximizing the distance metric between signal segments with different V-A labels. We tested EmoDSN on three different datasets collected in three different environments: desktop, mobile and HMD-based virtual reality, respectively. The results from our experiments show that EmoDSN achieves promising results for both one-dimension binary (high/low V-A, 1D-2 C) and two-dimensional 5-class (four quadrants of V- A space + neutral, 2D-5 C) classification. We get an averaged accuracy of 76.04, 76.62 and 57.62% for 1D-2 C valence, 1D-2 C arousal, and 2D-5 C, respectively, by using only 5 shots of training data. Our experiments show that EmoDSN can achieve better results if we select training samples from the changing points of emotion or the ending moments of video watching.**

***Index Terms*—Deep siamese network, emotion recognition, physiological signals, small data.**

## I. Introduction

A GROWING number of emotion recognition algorithms were developed in recent years [1]–[3] to model the temporal dynamics of emotion states of users. The accurate recognition of emotions while users consume different types of media content (e.g., videos, music, movies) can help content providers to better understand users' emotions towards the media content they provide and adjust it accordingly [4]. Unlike recognizing only one emotion label for a video clip (i.e., discrete emotion recognition), fine-grained (normally 0.5 s to 4 s according to prior emotion duration measures [1], [5], [6]) emotion recognition can capture the time-varying nature of human emotions [7]–[9]. Thus, the predictions are temporally more precise compared with discrete emotion recognition.

To model the temporal dynamics of emotions, physiological signals such as Electrodermal Activity (EDA), Blood Volume Pulse (BVP), Skin Temperature (TEMP) and Heart Rate (HR) are widely used by previous works [1], [2] as the input signals. These signals can represent the neural activities from both the autonomic nervous system (EDA and TEMP) and the cardiovascular system (BVP and HR). These activities provide sufficient information for V-A recognition [10], [11] according to *James-Lange* theory [12]. They are also easy to measure using unobtrusive and wearable sensing devices such as wristbands or smartwatches (e.g., Microsoft MS Band).

Previous works [1], [3], [13], [14] on fine-grained emotion recognition rely on large amounts of training data with fine-grained emotion labels. These labels are required to be collected in a fine level of granularity (normally the same or similar frequency as the input signal) to train the recognition algorithms [15]. To collect such fine-grained emotion labels, researchers either ask users themselves to label their emotions in real-time while watching videos [8], [16] or invite external annotators to label users' emotions segment-by-segment (e.g., using videos of users' facial expressions [17]) after watching the videos [17], [18]. However, it is challenging to collect large amounts of annotated signals using any of the methods. Asking users to momentarily self-report their emotions can incur more mental workload and result in user fatigue for longer durations (e.g., a two-hour film). For external annotators, at least three annotators are usually required to get a meaningful agreement between them (e.g., high Kappa score) [17], [19]. This requires extra labeling effort and is costly when annotating large amounts of signals. Thus, the experiments to collect large amounts of continuously annotated signals are time-consuming (require additional annotation time from users) and costly (hiring professional annotators is expensive).

The challenge of collecting large amounts of annotated signals has motivated researchers to explore Few-Shot Learning (FSL) algorithms [20] for emotion recognition. FSL algorithms are designed to converge on a small amount of training data and

Manuscript received 8 November 2021; revised 10 February 2022; accepted 25 March 2022. Date of publication 7 April 2022; date of current version 8 September 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yadong Mu. *(Corresponding author: Tianyi Zhang.)*

Abdallah El Ali is with the Distributed and Interactive Systems, Centrum Wiskunde and Informatica (CWI), 1098XG Amsterdam, The Netherlands (e-mail: abdallah.el.ali@cwi.nl).

Tianyi Zhang and Pablo Cesar are with the Distributed and Interactive Systems, Centrum Wiskunde and Informatica (CWI), 1098XG Amsterdam, The Netherlands, and also with the Multimedia Computing Group, Delft University of Technology, 2600AA Delft, The Netherlands (e-mail: tianyi@cwi.nl; p.s.cesar@cwi.nl).

Alan Hanjalic is with the Multimedia Computing Group, Delft University of Technology, 2600AA Delft, The Netherlands (e-mail: a.hanjalic@tudelft.nl).

Digital Object Identifier 10.1109/TMM.2022.3165715

provide relatively accurate prediction results. However, current FSL algorithms are geared towards discrete emotion recognition [21] using static data modalities such as images [22]. Thus, it is challenging to directly apply the existing FSL algorithms for fine-grained emotion recognition using physiological signals. First of all, there can be temporal mismatch between physiological signals and the fine-grained self-reports (i.e., the delay of annotation). Previous works [23]–[25] found that there are time delays between an emotional event and its annotation. The time of the delay ranges from 2 s to 4 s according to the experiments of Huang *et al.* [23]. Secondly, some fine-grained samples in the training set can be labeled incorrectly [1]. The mislabeled samples can be the result of a distraction of users when self-reporting their emotions momentarily or from a temporary failure of the system when collecting the labels. Both the reaction delay and mislabeled training samples can result in a mismatch between training samples and the corresponding ground truth labels. Since we only use few annotated samples for training, the mismatch can cause mis-convergence and overfitting for the recognition model. Previous works [23], [26] show that both the delay of annotation and mislabeled training samples can lower the accuracy if we directly build the recognition model between input signals and fine-grained emotion labels.

To overcome these challenges, this paper proposes a few-shot learning algorithm (*EmoDSN*[1]) for fine-grained emotion recognition on small data using physiological signals. *EmoDSN* is designed based on Deep Siamese Network (DSN), which can rapidly converge on a small amount of training data (typically $<$10 samples per class (i.e., 10 shot) [27]). It can provide recognition results at fine level of granularity (every 2 s) by maximizing the distance metric between signal segments with different emotion labels. To overcome the temporal mismatch between signals and emotion labels, we design an embedding network to automatically compensate for the delay of fine-grained emotion labels. To avoid overfitting caused by mislabeled samples, we also develop the distance fusion module which can merge the distance metrics learned from different training samples. This work makes the following contributions for multimedia community:

- We propose an end-to-end few-shot learning algorithm which can predict V-A in fine-level of granularity (2 s) using physiological signals trained by a small amount ($<$10 shot) of data. The algorithm can help researchers to understand the personalized experience of users watching videos by collecting only a small amount of data for training.
- We test our algorithm on three datasets (CASE [16], MERCA [8] and CEAP-360VR [28]) collected in three environments (desktop, mobile, and HMD-based Virtual Reality (VR)). Recognition results show good performance for both personalized binary (1D-2 C) and 5-class (2D-5 C) classification on all three datasets. we get an averaged accuracy of 76.04%, 76.62% and 57.62% for 1D-2 C valence, 1D-2 C arousal and 2D-5 C respectively by using 5 shot of training data. Our algorithm enables finding an optimal trade-off between recognition accuracy and collecting small amounts of continuously annotated physiological signals.
- We test state-of-the-art FSL algorithms [29]–[31] and compare their performance with *EmoDSN*. Results show that the recognition accuracy of *EmoDSN* outperforms other FSL algorithms. Our ablation study also shows that the embedding network (+11.86%) and distance fusion module (+22.32%) we design can significantly improve the accuracy.
- We run experiments to identify training samples from which temporal moments of video watching (e.g., begin, end and changing points [32]) can better represent the distribution of emotion labels and result in better recognition results. We find that the changing points of emotion annotation and the ending moments of video watching are better temporal moments for training samples (result in higher recognition accuracy) when only few annotated samples are available.

[1][Online]. Available: https://github.com/cwi-dis/EmoDSN

## II. Related Work

In this section, we first review the previous works on emotion recognition on small data and then narrow our scope to few-shot learning based emotion recognition.

### A. Emotion Recognition on Small Data

Fine-grained emotion recognition requires algorithms to predict multiple emotion states by relying on signals within one certain time interval. To train such recognition models, previous works [1], [3], [13], [14] need large amounts of data which are annotated in fine-level of granularity. Specifically, they usually require more than 90% of the annotated data in the datasets (e.g., CASE [16], RECOLA [17], K-EmoCon [19], MERCA [1]) to train an accurate recognition model. That means users themselves or external annotators have to continuously annotate 3 to 9 hours (e.g., CASE: 9.5 h, RECOLA: 3.4 hours, K-EmoCon: 5.3 hours, MERCA: 7.5 hours) to obtain an adequate amount of data for training. That requires large amounts of labeling effort for either external annotators or users themselves. Thus, it is challenging to collect large amounts of continuously annotated data for fine-grained emotion recognition.

To overcome this challenge, previous works have applied two kinds of methods to build recognition models with a small amount of training data. The first kind of method [33]–[37] builds a generative model such as Generative Adversarial Network (GAN) to generate artificial signals which obey the distribution of specific emotion categories. Then the recognition models are trained with the hybrid of synthetic and real signals. For example, Chen [33] *et al.* design a GAN model to generate ECG samples with the corresponding emotion labels. Their experiments show that the augmented dataset help to increase the accuracy by 5% compared with using only original data. Previous works on other physiological signals (i.e., Electroencephalography (EEG) [34], Electrooculography (EOG) [35], Blood Volume Pulse (BVP) [36], saccadic eye movement [37]) have also demonstrated that the augmented signals can promote the recognition accuracy by providing more data to train the recognition

model. However, to generate generalizable distributions for different emotion categories, the generative model itself also needs large amounts of signals with continuous annotation [38].

The second kind of method designs machine learning methods which can be trained by a small amount of ground truth labels. For example, Romeo *et al.* [39] implement four weakly-supervised learning algorithms to estimate fine-grained emotion states from post-stimuli emotion labels (i.e., the labels user annotate after each video watching). The methods they develop can identify which fine-grained signal segments (i.e. instances) can represent the post-stimuli valence and arousal [40]. Similar approach is used by Pei *et al.* [41] to model the temporal dynamics of emotional states. In their work, a weakly-supervised Bidirectional LSTM [42] is designed to predict fine-grained emotion labels according to the probability for that instance to predict the corresponding coarse labels. Although the weakly-supervised methods can predict fine-grained emotion labels with less amount of annotation, they can only identify the annotated (e.g., post-stimuli) emotion from the baseline emotion (e.g., neutral) and categorize all the remaining moments as part of the baseline. Thus, they can only predict two emotion states (i.e., the annotated emotion and neutral) in fine-level of granularity.

### *B. Few-Shot Learning Based Emotion Recognition*

Few-shot learning (FSL) is a kind of machine learning method which can learn a task from few (typically $<10$ samples per class [30], [31], [43]) annotated samples. Compared with weakly-supervised learning methods, FSL algorithms build direct mappings between fine-grained emotion labels and input signals, which can provide prediction with multiple emotion categories. FSL has been applied in previous works on emotion recognition using a variety of data modalities such as images [22]. To learn the representation of emotions using few annotated samples, researchers need to design different embedding networks for different data modalities. For example, Zhan *et al.* [22] design an affective structural embedding framework to predict the emotions of images. Their embedding network can learn an intermediate space which bridges the affective gap between low-level and high-level visual semantics.

For physiological signals, Jiang *et al.* [21] develop an FSL algorithm to recognize the level of stress using ECG, EDA and respiratory (RESP) signals. Their method, which is based on the Matching Network [29], achieves 80% accuracy trained by only 30% of the signals (i.e., 31.5 mins) in WESAD dataset [44]. Patane *et al.* [27] propose a siamese network based arousal recognition algorithm using ECG signals. Their algorithm obtains +21.5% accuracy increase compared to state-of-the-art machine learning algorithms trained with a subject-dependent model. Siamese network [45] is a kind of FSL algorithm which learns the difference between samples with different labels. Compared with other FSL algorithms (e.g., the Matching Network [29] used by Jiang *et al.*), Siamese network uses the pair-by-pair learning structure (learn the difference between two samples in two categories) instead of using the one-to-many learning structure (learn the difference between one sample and samples in other categories). It has been widely used for emotion recognition because of its simple and interpretable structure [46]. For example, Hayale *et al.* [46] use the Deep Siamese Neural (DSN) network [47] to recognize 6 basic emotions by facial expressions. For uni-dimensional signals, DSN is also used by Feng *et al.* [48] to predict low/medium/high arousal using speech signals. They obtain 43.4% accuracy trained with a subject-dependent model.

Although the previous works above provide useful insights on FSL or DSN based emotion recognition, they only recognize the overall emotion of an event (e.g., one video watching) instead of the fine-grained emotion responses. Our work aims to extend few-shot learning algorithms for emotion recognition with fine-level of granularity.

## III. DSN Based Emotion Recognition

In this section, we propose an Emotion recognition algorithm based on Deep Siamese Network (*EmoDSN*) to discriminate fine-grained physiological signal segments (i.e., samples) with different emotion labels. *EmoDSN* learns the difference between samples instead of building the precise mapping between samples and emotion labels. Thus it can converge with only few annotated samples as training data. In the training stage, $n$ samples are used for training. The influences of different temporal moments of training samples are discussed in Section VI-B. *EmoDSN* contains four parts: (1) **Pre-processing:** the obtained physiological signals are firstly pre-processed using different filters to remove the noise and artifacts in signals. (2) **Embedding Network**: the pre-processed signals are then fed into an embedding network to learn embeddings representing the difference of samples between emotion labels. (3) **Siamese Learning**: the embeddings are learned based on the siamese structure. The output of siamese learning is a distance metric which can represent the probability that the two input samples belong to the same emotion label. After the network is learned, the embedding for each training sample will also be generated. In the prediction stage, the pairwise distance metrics between testing and training samples are fused by the (4) **Distance Fusion** module to obtain the probability of the testing samples corresponding to different emotion labels. The testing samples are predicted as the emotion label with the highest probability. Below we provide a detailed description of *EmoDSN*.

### *A. Pre-Processing*

The physiological signals are first pre-processed by different filters to remove the noises and artifacts. We follow the pre-processing procedures which are widely used in previous works [10]. For EDA signals, a low pass filter with a 2 Hz cutoff frequency is used to remove noise [49]. For the BVP signals, a 4-order butterworth bandpass filter with cutoff frequencies [30, 200] Hz is implemented to eliminate the bursts [50]. For TEMP signals, we use an elliptic band-pass filter with cutoff frequencies [0.005, 0.1] [51]. To decrease measurement bias in different sessions (i.e., each subject under each video stimulus), all signals for each session are normalized to [0,1] using Min-Max scaling normalization.

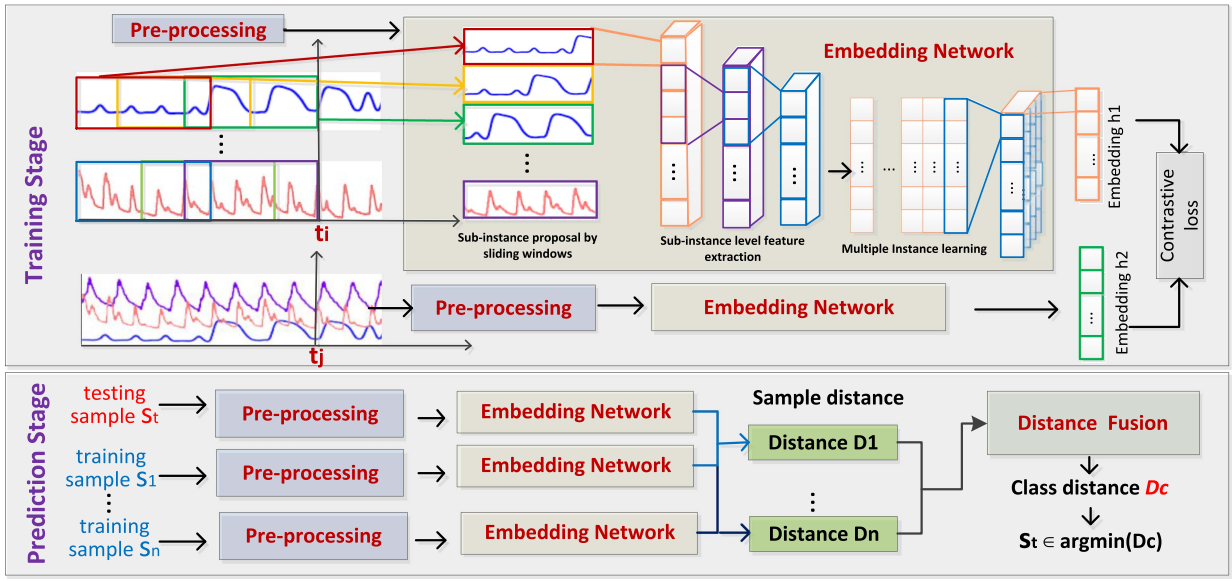


Fig. 1. The architecture of proposed EmoDSN.

### B. Embedding Network

The purpose of the embedding network is to automatically extract features and learn latent vectors of samples to represent the difference of samples between emotion labels. Previous works [1], [52] on embedding networks for physiological signals usually use the fine-grained segments of signals (i.e., samples) as the input. However, samples can be misaligned with the fine-grained emotion labels due to the reaction delay of continuous self-reporting. When continuously annotating emotions towards videos, users first process the stimuli using their senses (at $t_s$) and then react to these changes (at $t_s + t_d$) which leads to a reaction delay ($t_d$). Thus, the sample at $t_s$ actually corresponds with emotion label at $t_s + t_d$. To address the problem of reaction delay, we use sliding windows to propose multiple signal segments (sub-instance) with different delays. Then, the embeddings of these sub-instances are learned using a weakly-supervised multiple instance learning network. Below, we describe the implementation details of each module in the embedding network.

*1) Sub-Instance Proposal:* Suppose $S = \{s_n\}_{n=1}^{N}, s_n \in R^{L\times C}$ is a set of physiological signals with the number of channels $C$ and the segmentation length $L$. For each sample $s_n$, there is a corresponding emotion label $l_m$. To generate embeddings which consider the delay of reaction, we reconstruct $s_n$ with multiple sub-instances $s'_n = [s_{n1}, s_{n2} \dots s_{nK}]^T$, where $s_{nk}$ is a sub-instance (i.e., each row of sample $s'_n$) with the delay of $t_k$. We use sliding windows with the window length $L$ and stride $k$ to generate the sub-instances. After that, the input of the algorithm become $S' = \{s'_n\}_{n=1}^{N}, s'_n \in R^{K\times L\times C}$, where $K$ is the number of sub-instances for each $s'_n$.

*2) Sub-Instance Level Feature Extraction:* The features are extracted from each sub-instance $s_{nk}$ independently, which means the feature extraction layers will not influence the independence between each sub-instance (no features are extracted from multiple sub-instances). The independent feature extraction guarantees that each sub-instance has a unique instance gain after the embedding network. The instance gains can help us understand the duration of delay with which the network can best discriminate signal segments with different emotion labels.

The features for each sub-instance are extracted using a 3-layer (kernal size: $L/2+1-L/4+1-L/8+1$, channels: 4-8-16) 1D-CNN [53]. We use a shallow structure (three layers) instead of deep to avoid overfitting since each sub-instance does not contain much information. We use large (i.e., equals to half of the sub-instance length) convolutional kernels in the shallow layer of the network. Large convolutional kernels have a large receptive field across different sampling points in one sub-instance thus can result in better recognition accuracy [54]. However, the local information can also be omitted by large kernels and result in the difficulty for the network to converge [55]. Thus, we follow a classical strategy that gradually increases the number of kernels and decreases the size of them when the network goes deeper [56]. After sub-instance feature extraction, the $S' = \{s'_n\}_{n=1}^{N}$ is mapped to the feature vectors $F = \{f_n\}_{n=1}^{N}, f_n \in R^{K\times L\times E}$, where $E = 16$ is the dimension of feature vectors.

*3) Multiple Instance Learning:* The purpose of multiple instance learning module is to 1) merge the features learned in sub-instances to generate embeddings and 2) assign each sub-instance a instance gain representing the weights of sub-instances for discriminating samples with different emotion labels. The instance gains for all the sub-instances construct the embeddings for the sample. Here we use a weakly-supervised multiple instance learning architecture which is shown in Fig. 2. Multiple instance learning can map the feature vectors of sub-instances to the probability for that sub-instance to specific task (in our case, discriminating between emotion labels). Thus, it can promote the interpretability of our algorithm by helping us understand with how much delay (sub-instances with high probability) the signal segment can better predict emotions.

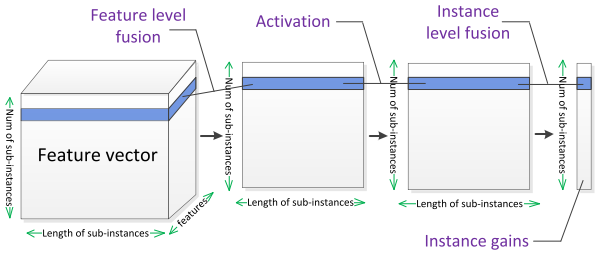


Fig. 2. The diagram for multiple instance learning module.

The feature vectors obtained from the previous module are first input into a feature level fusion module using uni-dimension convolution. The convolution is conducted on the dimension of $E$ to merge the features from different signal channels. After that, the merged features are activated by a Rectified Linear Unit (ReLU) function. Another uni-dimension convolution is implemented on the dimension of $L$ to fuse features for different sampling points inside each sub-instance. At last, we activate the results from previous modules with a softmax function. The purpose of the softmax activation is to (a) normalize the instance gains in the range from 0 to 1 and (b) make the network easier to calculate the gradient for back-propagation. After the multiple instance learning module, the feature vector $f_i = \{f_{nk}\}_{k=1}^{K}, f_{nk} \in R^{L\times E}$ is mapped into instance gain $g_n = \{g_{nk}\}_{k=1}^{K}, g_n \in R^1$. At last, the embedding of one sample $h = [g_1, g_2, \ldots, g_K]$, where $K$ is the number of sub-instances for one sample.

### C. Siamese Learning

The purpose of using the siamese learning network is to learn a distance metric which can discriminate samples with different emotion labels. Specifically, for sample $s_i$ and $s_j$, which are two signal segments for $t_i$ and $t_j$, the siamese learning network learn a distance metric $D$ with the target of $D = 0$ if they are with the same emotion label and $D = 1$ if they are with the different emotion labels. To train the network, we first construct two embedding networks with shared weights. The two embeddings $h_i$ and $h_j$ generated from the network are trained by contrastive Loss:

$$L_{contrast} = (1 - \mathrm{Y})\frac{1}{2}(\mathrm{D_w})^2 + \mathrm{Y}\frac{1}{2}\max(0, 1 - \mathrm{D_w})^2 \quad (1)$$

where Y equals 0 or 1 for $s_i$ and $s_j$ have the same or different emotion labels respectively. $D_w$ is the euclidean distance for $h_i$ and $h_j$. We also tested the cosine distance metric which is also widely used for other siamese networks. However, our network cannot converge using cosine metric. The contractive loss encourages the network to learn embeddings to place samples with the same labels close to each other while distancing the samples with different emotion labels in the embedding space. The siamese learning network is trained with the *RMSprop* [57] optimizer because it can automatically adjust the learning rate for faster convergence.

### D. Distance Fusion

In the prediction stage, when a new sample $s_t$ at time $t$ comes, we can obtain the pairwise distance metric $D = \{D_n\}_{n=1}^{N}$ by calculating euclidean distance between $s_t$ and all training samples $\{s_n\}_{n=1}^{N}$ using their embeddings. The distance metric $D$ can also be used to represent the probability of $s_t \in l_m$ if the emotion label of $s_{nm}$ is available:

$$P(s_t \in l_m | s_{nm} \in l_m) = 1 - D \quad (2)$$

where $P(s_t \in l_m | s_{nm} \in l_m)$ represents the probability that $s_t$ corresponds to the emotion label $l_m$ under the condition of $s_{nm} \in l_m$. Previous works [22], [29], [31] on few-shot learning simply average $D$ with the same emotion labels and predict $s_t$ as the emotion label with the closet distance (or greatest possibility). However, the hypothesis of averaging the distances is that the labels for all training samples are correct:

$$P(s_t \in l_m) = \sum_{m=1}^{M} P(s_t \in l_m | s_{nm} \in l_m) \cdot P(s_{nm} \in l_m) \quad (3)$$

From (3) we can conclude that if all $P(s_{nm} \in l_m) = 1$, $1 - P(s_t \in l_m)$ equals to the average of $D$. However, the fine-grained self-reports, which are used as the labels for training, can be mismatched with the physiological signals. Thus, some samples in the training set can be labeled incorrectly. This problem is not that severe when we use large amounts of samples for training. However, when we only use few annotated samples, one or multiple mislabeled samples can significantly lower the model accuracy.

To solve this problem, we propose the **Distance Fusion** module based on Bayesian Fusion to estimate $P(s_{mn} \in l_m)$. Suppose there are $N$ training samples which are annotated as $M$ emotion labels, $N = \{N_m\}_{m=1}^{M}$ are the numbers of training samples with $M$ emotion labels, respectively. $N_m$ is the number of training samples labeled as $l_m$. $s_{mn}$ represents training sample $n$ annotated as emotion label $l_m$. The probability of $s_{mn} \in l_m$ can be estimated by:

$$P(s_{mn} \in l_m) = 1 - \frac{1}{2}\left[\frac{1}{N_m}\sum_{k=1}^{N_m} D_{mk} - \frac{1}{M-1}\sum_{i=1}^{M, i\neq m}\left(\frac{1}{N_i}\sum_{j=1}^{N_i} D_{ij}\right)\right] \quad (4)$$

where $D_{ij}$ represent the distance between training sample $s_i$ and $s_j$. The first and second $\Sigma$ terms of (4) represent the probability of $s_{mn}$ similar to the training samples with the same and different emotion labels of $s_{mn}$ respectively. If $s_{mn}$ is similar to the samples with the same label and dissimilar with the samples with different labels, the probability of $s_{mn} \in l_m$ is high.

After we obtain all $P(s_{mn} \in l_m)$ for $N_m$ samples labeled as $l_m$, we can calculate $P(s_t \in l_m)$ by (3). At last, we predict $s_t$ corresponds to the emotion label with the highest probability:

$$l_t = \arg\max_m (P(s_t \in l_m)) \quad (5)$$

where $l_t$ is the predicted emotion label for the $s_t$.

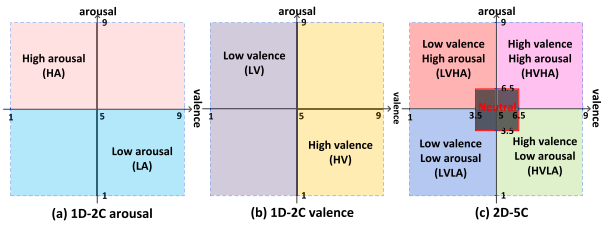


Fig. 3. Graphical illustration of discretized emotion categories.

TABLE I
THE PERFORMANCE OF *EmoDSN* TRAINED WITH 5-SHOT

| 5-shot | CASE | | MERCA | | CEAP-360VR | |
|---|---|---|---|---|---|---|
| | acc | m-f1 | acc | m-f1 | acc | m-f1 |
| **1D-2C-valence** | 77.30% | 0.722 | **78.95%** | **0.765** | 71.86% | 0.695 |
| **1D-2C-arousal** | **77.72%** | 0.709 | 77.18% | **0.764** | 74.95% | 0.729 |
| **2D-5C** | **58.77%** | 0.482 | 56.08% | 0.508 | 56.93% | **0.512** |

## IV. DATASETS

We test *EmoDSN* on three datasets: *CASE* [16], *MERCA* [8] and *CEAP-360VR* [28] which are collected in three environments: desktop, mobile and VR, respectively. The problem *EmoDSN* focuses on is recognizing valence and arousal (V-A) in fine level of granularity. Thus, we choose datasets with fine-grained V-A self-reports as ground truth labels for validating the performance of *EmoDSN*. We evaluate *EmoDSN* on datasets collected in three different environments to test whether it can be generalized to different scenarios. We also test *EmoDSN* by signals collected using golden standard (CASE) and wearable (MERCA and CEAP-360VR) devices to test whether *EmoDSN* can generalize to different types of physiological sensors.

## V. EXPERIMENTS AND RESULTS

### A. Implementation Details

To implement a fair evaluation among the three datasets, we process the physiological signals to be as similar as possible before inputting them to *EmoDSN*. Since the three datasets have different sampling rates, we interpolate the signals in MERCA and CEAP-360VR to 50 Hz using linear interpretation [58]. We choose linear interpolation because it is the simplest interpolation method which will not change the distribution of the signals. For the CASE dataset, the signals are down-sampled to 50 Hz by decimation down-sampling [59]. The HRs signals of CASE are extracted from ECG signals using *heartpy* library [60]. We use the mean V-A value of 2-second [1] as the labels for training and testing the algorithm. The window length $L$ and stride $k$ for the sub-instance proposal are 2 s and 0.5 s respectively according to previous research [1], [39]. For each timestamp $t$, we move the sliding window 12 times to cover the annotation delay for maximum 10 s. The amount of time of annotation delay is discussed in Section VI-A.

We evaluate *EmoDSN* by two tasks: the one-dimensional two-class (1D-2 C) classification [61] and the two-dimensional 5-class (2D-5 C) classification [62], which are widely used as the tasks for evaluating emotion recognition algorithms using physiological signals. We follow the standard labeling schemes from previous works [61], [62] to map continuous values of V-A to discretized emotion categories. The graphical illustration of this operation is listed in Fig. 3.

For the training procedure, we train user-specific models for all the users in three datasets. We follow the standard procedure of testing few-shot learning algorithms with continuous signals [43]. We randomly sample $N$ (i.e., shot) sampling points in each emotion category as training samples from one user and use the rest of the samples for testing. The results reported in this section are the average results for all users. We also tried to train user-independent models which use only few annotated samples from one user and test the model on other users. However, due to the high inter-subject variability that affects the physiological signals, building user-independent emotion recognition model is still challenging even using large amounts of annotated data [39], [63]. In this study, we use only few annotated data for training. User-independent models did not achieve satisfactory performance (accuracy not above chance level) for all three datasets and thus the result was not reported in this study.

### B. Classification Results

We use accuracy (acc) and macro-F1 score (m-f1) to evaluate the performance of our algorithm. The accuracy represents the percentage of correct predictions. The macro-F1 score is the mean of precision and recall for each label. We use macro-F1 score instead of weighted and binary F1-score to take into account label imbalance. Compared with accuracy, the macro-F1 score can provide more objective evaluation results by taking into account how the data are distributed.

The performance of *EmoDSN* trained with 5-shot is shown in Table I. *EmoDSN* can obtain up to 70% for 1D-2 C and 56% for 2D-5 C, which are much higher than chance level (shown in Fig. 6). The results are obtained by training with only 5-shot (10 seconds of sampling points for each emotion category). That demonstrates that *EmoDSN* can converge and obtain accurate fine-grained emotion recognition with only few annotated samples.

### C. Results for Different Emotion Categories

The confusion matrices for 2D-5 C are shown in Fig. 4. We only show the confusion matrices for 2D-5 C because it contains classification results for more emotion categories. From the confusion matrices we can see that *EmoDSN* performs well on discriminating the neutral and non-neutral samples. Almost all neutral samples are predicted as neutral for the three datasets. *EmoDSN* also performs well on discriminating samples with high valence. An averaged acc of 78.6% is obtained by *EmoDSN* when discriminating high/low arousal under the condition of high valence. However, the performance on discriminating samples with low valence is not as good as high valence. More than 20% of the LVHA samples are categorized as LVLA on average of the three datasets.

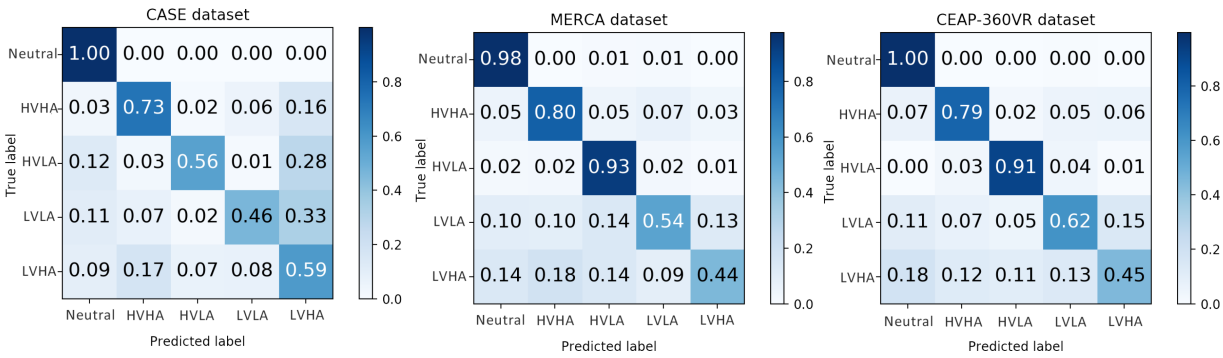


Fig. 4. The confusion matrices for 2D-5 C trained by 5-shot.

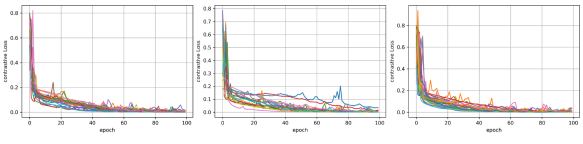

Fig. 5. The training losses (5-shot, 2D-5 C) of EmoDSN on CASE (left), MERCA (middle) and CEAP-360VR (right), each curve represents the training loss for the user-specific model trained on one user.

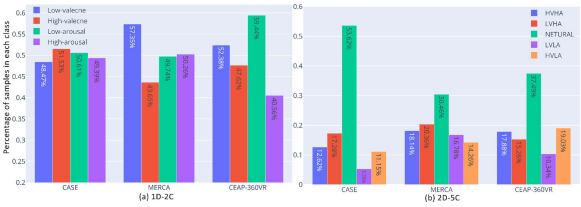


Fig. 6. Percentage of samples in different emotion classes.

The reason is the class imbalance for different emotion categories. As shown in Fig. 6(a), for 2D-5 C, the neutral class has more than 30% of samples for all three datasets. The most imbalanced dataset is CASE, which contains more than 50% of samples with neutral labels. We also find the LVLA and LVHA classes have comparatively fewer samples (16.48%, 31.04% and 29.37% for CASE, MERCA and CEAP-360VR respectively). That is why discriminating different levels of arousal is more challenging under the condition of low valence. However, for 1D-2 C, the high/low V-A classes are balanced. We do not find any classes with less than 40% of all samples. Thus, the performance for 1D-2 C is comparatively balanced: the m-F1 score is 3.6% lower than the acc on average of the three datasets. For 2D-5 C, the m-F1 is 7.4% lower than the acc on average of the three datasets.

However, even taking the class imbalance into consideration, the acc obtained by *EmoDSN* is still higher than the chance level. For CASE, MERCA and CEAP-360VR respectively, the accuracies are 19.07%, 24.67% and 17.75% higher than the percentage of samples in the class with the most samples (i.e., the chance level). The results show that although *EmoDSN* provides relatively imbalanced precision and recall for different V-A categories, it does not overfit into one specific V-A category and can still provide accurate predictions.

### *D. Results for Different Datasets and Subjects*

For the comparison between different datasets, our method performs best on CASE dataset (up to 76% and 58% acc for 1D-2 C and 2D-5 C respectively). The acc of 1D-2 C on MERCA is similar to CASE but the 2D-5 C acc on MERCA is 2.69% lower. Both the accuracies for 1D-2 C and 2D-5 C on CEAP-360VR are lower than the accuracies on CASE for 3.35% on average. We speculate that the different accuracies of *EmoDSN* on three datasets is a result of the different experimental environments. The data collection experiment of CASE was conducted in an indoor laboratory environment, which contains less interference and noise (e.g., environment noise, user movement, sensor detachment). Thus, the signals from CASE contain less noise and artifacts caused by both the users themselves and the outside environment. The results indicate that the mobile (MERCA) and VR environments (CEAP-360VR) are more challenging for fine-grained emotion recognition compared with a laboratory-based desktop (CASE) environment. However, the maximum difference in acc between the three datasets is less than 7%, which shows that our algorithm does not overfit on one specific dataset. The test results on different datasets show good generalizability of *EmoDSN* among different environments (desktop, mobile and VR).

For the comparison between different subjects, Fig. 7 shows the acc for each individual subject of three datasets. From Fig. 7 we can find variability of acc between different individuals: the average SD for 1D-2 C valence, arousal and 2D-5 C are 10.43%, 11.31% and 6.11% respectively. Our model achieves up to the chance level (the percentage of samples in the class with the most samples) accuracies for 86.05%, 85.57% and 82.37% of the subjects for 1D-2 C valence, 1D-2 C arousal and 2D-5 C respectively. For the subjects which our algorithm does not achieve above the chance level accuracies, we find the annotations of their data are highly imbalanced (i.e., subject annotates a high percentage of neutral emotion). For example, subject 6 in the CEAP-360VR dataset annotated 72.35% of his or her emotion as neutral when watching videos. The average percentage of neutral annotations for these subjects is 28.41% higher than the subjects whose accuracies are above the chance level. Although recognition accuracies from some of the subjects are low because of class imbalance, our model still achieves above the chance level acc for more than 80% of the subjects. The balanced

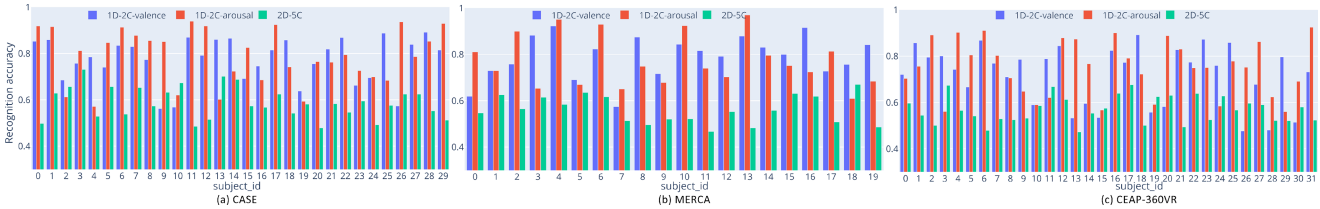


Fig. 7. The recognition acc for individual subject of CASE, MERCA and CEAP-360VR.

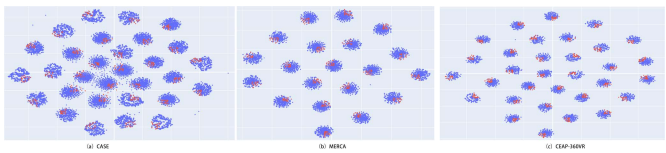

Fig. 8. The visualization of embedded features using t-SNE. Red and blue points denote the train and test samples, respectively. Best viewed in color.

performance on different subjects shows good generalizability of *EmoDSN* among different subjects.

### *E. Visualization of the Embeddings*

To visualize the joint sample distribution of the training/testing set, we use T-distributed stochastic neighbor embedding (t-SNE) to reduce the dimension of the embeddings to 2D. It is widely used by previous works [64], [65] of few-shot learning approaches for visualizing the training/testing set. From Fig. 7 we can see that the embedding trained by *EmoDSN* constructs the compact clusters of the training features close to the testing features (average purity score = 0.685, 0.657 and 0.614 for CASE, MERCA and CEAP-360VR respectively). The close temporal position between training/testing samples indicates that the learned embeddings can represent the joint distributions between few-shot training samples and the remaining test samples for emotion classification. Previous works [64], [66] show that the closer the training/testing sets are, the easier the classification network can complete the learning task. Our visualization results demonstrate the effectiveness of the embedding network we designed for the Deep Siamese Network.

### *F. Comparison With Baseline Methods*

*1) Implementation Details:* To compare the performance of *EmoDSN* with state-of-the-art emotion recognition methods, we choose two kinds of baselines: classic FSL networks (i.e., Matching network (MN) [29], Prototype network (PN) [30], Deep Siamese Network (DSN) [47], Relation Network(RN) [31] and Model Agnostic Meta Learning (MAML) [43]) and networks designed for physiological-signal-based emotion recognition (HetEmotionNet (HetNet) [67] and SFENet [68]). We choose the five FLS baselines because they are widely used by previous works for emotion recognition using similar data modalities (i.e., uni-dimensional data modalities such as speech [48] and physiological signals [21]). To implement a fair comparison, we fine-tune the structure of these methods to make them have the same embedding network we designed in Section III-B. Thus, the difference between each method is only the learning structure instead of the embedding network for feature extraction. We also use the same optimizer and learning rate (lr = 0.001) as *EmoDSN* to train all four few-shot learning algorithms. For HetNet, we construct the spatial-temporal and spatial-spectral graph (by DE features) and train them using the same graph recurrent neural network. Since the folding approach used by SFENet is based on the spatial distribution of EEG electrodes, we cannot use it for other physiological signals. Thus, we only use the 3D-CNN and ensemble learning designed in SFENet for comparison. We train the above algorithms with one, five and ten shot to compare their performance trained by different amounts of annotated samples. To test the stability of each algorithm, we run all the experiments 5 times [69], [70] and report the mean and SD of the accuracies.

*2) Accuracy Comparison:* Table II shows the results of the comparison. We observe that the gradient cannot descent (losses remain constant) when training the MN, PN and RN with 1-shot and 5-shot for 2D-5 C. As shown in Fig. 5, this problem does not occur when we use DSN: the losses descend rapidly after a few epochs (<10) for all personalized models in three datasets. The performance of MAML is better (acc is 8.89% and 6.94% higher for 1D-2 C and 2D-5 C respectively) than other FSL methods. However, the acc increase for MAML is not as significant as other FSL methods: when the number of training samples increases from 1-shot to 5-shot, the acc increase 3.04% and 6.94% on average for 1D-2 C and 2D-5 C respectively. The other FSL methods however, increases 9.22% and 10.76% for 1D-2 C and 2D-5 C respectively.

For the two fully supervised learning methods, we find a similar problem with MN, PN, and RN that for 2D-5 C, the gradient cannot descent (losses remain constant) for 1-shot and 5-shot. Their average acc for 10-shot is also 5.89% lower than DSN. We also find the problem of overfitting for them when trained with 10-shot: the training acc increases rapidly over 90% after 5 epochs but the testing acc does not increase. The results demonstrate that the fully-supervised learning methods cannot achieve good performance when only a limited amount of data are used for training.

In general, the performance of *EmoDSN* is better than both the state-of-the-art FSL algorithms and fully-supervised algorithms. To compare the performance difference between *EmoDSN* and baseline methods, we follow the previous work of Kumar *et al.* [71] which use *Z-test* and *Chi-square* test to compare the classification accuracies. For both the *Z-test*

TABLE II
COMPARISON BETWEEN FEW-SHOT LEARNING METHODS

| Dataset | Methods | 1-shot accuracy | | | 5-shot accuracy | | | 10-shot accuracy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1D-2C valence | 1D-2C arousal | 2D-5C | 1D-2C valence | 1D-2C arousal | 2D-5C | 1D-2C valence | 1D-2C arousal | 2D-5C |
| **CASE** | MN [29] | .364(.123) | .381(.132) | .202(.131) | .446(.145) | .501(.125) | .303(.096) | .481(.199) | .554(.174) | .332(.166) |
| | PN [30] | .311(.129) | .317(.131) | .194(.126) | .467(.124) | .488(.140) | .317(.095) | .453(.184) | .568(.191) | .336(.131) |
| | RN [31] | .335(.131) | .355(.152) | .183(.132) | .381(.146) | .392(.127) | .224(.105) | .371(.182) | .366(.184) | .281(.150) |
| | DSN [47] | .409(.156) | .338(.132) | .368(.154) | .510(.101) | .478(.093) | .414(.097) | .563(.163) | .587(.176) | .482(.182) |
| | MAML [43] | .489(.142) | .476(.150) | .263(.151) | .495(.128) | .507(.166) | .361(.139) | .519(.140) | .526(.135) | .403(.146) |
| | HetNet [67] | .353(.066) | .374(.076) | .233(.088) | .465(.073) | .533(.058) | .364(.070) | .502(.054) | .532(.050) | .425(.077) |
| | SFENet [68] | .391(.086) | .396(.079) | .268(**.053**) | .411(.083) | .422(.078) | .275(.057) | .434(**.030**) | .426(**.025**) | .350(.074) |
| | **EmoDSN** | **.668(.062)** | **.654(.054)** | **.453**(.081) | **.778(.021)** | **.769(.044)** | **.583(.045)** | **.782**(.080) | **.778**(.096) | **.586(.110)** |
| **MERCA** | MN [29] | .371(.140) | .362(.133) | .192(.156) | .457(.136) | .524(.157) | .366(.123) | .455(.197) | .511(.178) | .356(.172) |
| | PN [30] | .283(.125) | .303(.142) | .211(.123) | .416(.145) | .511(.138) | .365(.109) | .435(.188) | .546(.178) | .385(.132) |
| | RN [31] | .365(.144) | .381(.138) | .185(.150) | .402(.130) | .451(.132) | .264(.120) | .369(.160) | .384(.192) | .237(.175) |
| | DSN [47] | .370(.127) | .405(.134) | .383(.124) | .429(.106) | .562(.090) | .446(.104) | .482(.184) | .602(.158) | .466(.173) |
| | MAML [43] | .517(.180) | .497(.167) | .393(.168) | .571(.146) | .582(.146) | .404(.153) | .568(.132) | .587(.098) | .421(.139) |
| | HetNet [67] | .400(.073) | .421(**.054**) | .279(.085) | .542(.082) | .536(.057) | .426(.086) | .557(.033) | .580(**.047**) | .435(**.088**) |
| | SFENet [68] | .404(.084) | .410(.080) | .247(**.074**) | .393(.083) | .397(.060) | .312(.053) | .448(**.026**) | .443(.054) | .337(.103) |
| | **EmoDSN** | **.683(.053)** | **.633**(.064) | **.432**(.090) | **.799(.033)** | **.763(.033)** | **.558**(.033) | **.802**(.065) | **.766**(.075) | **.553(.091)** |
| **CEAP 360VR** | MN [29] | .394(.138) | .423(.154) | .176(.168) | .411(.123) | .437(.123) | .357(.101) | .396(.196) | .502(.199) | .326(.178) |
| | PN [30] | .292(.125) | .336(.133) | .186(.156) | .407(.147) | .450(.155) | .345(.112) | .402(.185) | .521(.159) | .312(.136) |
| | RN [31] | .385(.141) | .386(.149) | .185(.150) | .396(.135) | .403(.136) | .271(.109) | .413(.160) | .425(.163) | .276(.126) |
| | DSN [47] | .401(.154) | .381(.135) | .358(.138) | .434(.094) | .507(.097) | .433(.082) | .473(.157) | .554(.194) | .446(.188) |
| | MAML [43] | .481(.144) | .486(.139) | .326(.138) | .496(.174) | .499(.124) | .424(.184) | .514(.130) | .521(.128) | .421(.107) |
| | HetNet [67] | .386(.058) | .360(**.080**) | .314(**.056**) | .538(.077) | .526(.053) | .436(.055) | .546(.049) | .558(.057) | .467(.081) |
| | SFENet [68] | .406(**.057**) | .401(.087) | .261(.064) | .411(.084) | .425(.056) | .315(.072) | .443(**.027**) | .435(**.030**) | .333(.106) |
| | **EmoDSN** | **.598**(.062) | **.625**(.084) | **.487**(.097) | **.720(.044)** | **.745(.035)** | **.561(.029)** | **.725**(.066) | **.742**(.077) | **.554(.073)** |

and *Chi-square* test, we found significant differences (all p<0.01) between EmoDSN and MN ($Z = 14.21, \chi^2 = 5.48$), PN ($Z = 14.14, \chi^2 = 6.59$), RN ($Z = 21.04, \chi^2 = 8.98$), DSN ($Z = 14.49, \chi^2 = 2.99$), MAML ($Z = 12.36, \chi^2 = 2.17$), HetNet ($Z = 11.69, \chi^2 = 3.06$) and SFENet ($Z = 23.74, \chi^2 = 5.86$). The statistical analysis shows a significant difference between the performance of *EmoDSN* and other baseline methods.

*3) Stability Comparison:* The stability of 5 FSL methods (i.e., MN, RN, PN, DSN, MAML) is lower than the two supervised learning algorithms (i.e., HetNet and SFENet): the SD for the 5 experiments is 7.81% higher. When the number of training samples increases to 10-shot, the SD difference between FSL and fully-supervised learning methods also increases accordingly (on average 10.75% for 10-shot). FSL algorithms learn the difference (MN, RN, PN, DSN) or train a meta learner (MAML) between training samples instead of learning the exact mapping between samples and labels. Thus, their performance depends on the quality of training samples, which leads to instability if we consider all training samples to be correctly labeled [72]. The fully-supervised learning methods however, optimize the classifier among all training samples. Thus, they converge on a worse (i.e., low acc) but comparatively stable model if only few samples are used for training. The results are in line with our conclusion in ablation study that we cannot get stable and accurate recognition results if we assume all $P(s_{nm} \in l_m)$ are to be 1.

TABLE III
ABLATION STUDY (ACC (SD)) FOR VANILLA SIAMESE (VS), EMBEDDING NETWORK (EN) AND DISTANCE FUSION (DF)

| | Dataset | VS | VS+EN | EmoDSN VS+EN+DF |
|---|---|---|---|---|
| **1D-2C valence** | **CASE** | .412(.119) | .480(.101) | .773(.021) |
| | **MERCA** | .337(.134) | .558(.106) | .789(.033) |
| | **CEAP-360VR** | .413(.145) | .510(.093) | .718(.043) |
| **1D-2C arousal** | **CASE** | .354(.135) | .501(.092) | .777(.044) |
| | **MERCA** | .365(.131) | .437(.089) | .772(.033) |
| | **CEAP-360VR** | .320(.147) | .440(.097) | .749(.035) |
| **2D-5C** | **CASE** | .314(.162) | .406(.097) | .587(.045) |
| | **MERCA** | .293(.159) | .443(.104) | .561(.034) |
| | **CEAP-360VR** | .322(.162) | .423(.080) | .569(.029) |

### G. Ablation Study

*1) Implementation Details:* We conduct an ablation study to verify the effectiveness of each component in *EmoDSN*. We begin with only using the Vanilla Siamese (VS) structure to train the network. The VS structure directly uses the raw signal segments without passing them through the embedding network. Then we test the performance of combining the VS with the Embedding Network (EN) described in Section III-B. For the two above experiments, instead of using Distance Fusion (DF), we follow the traditional strategy of few-shot learning algorithms: average the distances with the same emotion labels and predict

the samples as the emotion label with the closest distance. Finally, we replace the simple averaging with the DF described in Section III-D for the complete *EmoDSN*. To test the stability of *EmoDSN*, we also repeat the experiments 5 times [69], [70] and report the mean and SD of acc.

*2) Accuracy Comparison:* From the results (shown in Table III) we can see both EN and DF contribute to the classification tasks. The EN benefits the classification tasks by extracting deep features and taking reaction delay into consideration. Thus, the accuracies increase 11.85% on average after combining EN to VS. We also observe a significant increase of accuracies (more than 20% for 1D-2 C and 10% for 2D-5 C) after adding the distance fusion module. This finding demonstrates that simply averaging the distances from different shot is not suitable for fine-grained emotion recognition using physiological signals. It necessitates considering the probability that some mislabelled training samples can significantly lower the model accuracy. In conclusion, the observations above demonstrate the effectiveness of the components in the proposed algorithm.

*3) Stability Comparison:* For the comparison of SD, we find both VS and VS-EN have relatively unstable performance: the average SD is 14.39% and 9.58% for VS and VS-ED respectively. Adding DF however, can improve the stability of the network by decreasing the SD to 1.6%. When randomly selecting only few training samples, some samples with low-confidence annotation will affect the performance of the network. If all $P(s_{nm} \in l_m)$ are assumed to be 1, the network is unstable because the performance is related to the quality of labels selected for training. However, DF modules can decrease the instability by assigning less confident samples lower weights for classification. The results demonstrate the necessity and effectiveness of adding DF into *EmoDSN*.

### H. Effectiveness of DF Module

To further clarify the effectiveness of the distance fusion (DF) module for the wrong labels, we use it to identify potentially wrong labels and correct them when classifying emotions. Specifically, we first calculate the $P(s_{nm} \in l_m)$ for all training samples using (4). $P(s_{nm} \in l_m)$ represents the probability of training sample $s_{nm}$ corresponds to the emotion label $l_m$. If the probability is lower than 0.5, we assume the sample is mislabeled and correct it. Here we only run the experiments for 1D-2 C because we cannot estimate the correct label of $s_{nm}$ for multi-class classification if $P(s_{nm} \in l_m)$ is low. For multi-class classification, if we know $P(s_{nm} \in l_m)$ <0.5: we do not know which $i$ can satisfy $P(s_{nm} \in l_{i,i\neq m})$. However, for binary classification, since $P(s_{nm} \in l_m)$+$P(s_{nm} \in l_{i,i\neq m}) = 1$, if $P(s_{nm} \in l_m)$<0.5 we can easily know $P(s_{nm} \in l_{i,i\neq m})$>0.5. Thus, the mislabeled samples are corrected as the label opposite to its original annotation. Then we average the distance (D) with the same emotion labels after the label correction and predict the testing sample as the emotion label with the greatest possibility. Then we compare the recognition accuracies among the network a) without the Correction of Labels (no-CL), b) with the Correction of Labels (CL) and c) with the DF module. To ensure the stability of the experiment, we run the experiment 5 times [69], [70] and report the average acc and the SD of 5 experiments.

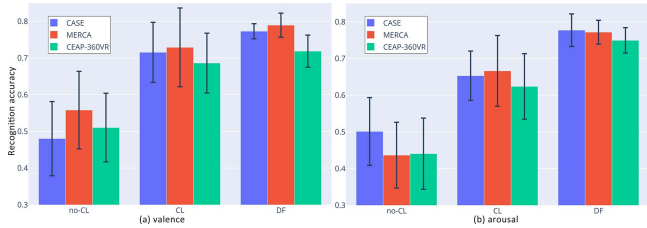


Fig. 9. Recognition accuracies among the network a) without the Correction of Labels (no-CL), b) with the Correction of Labels (CL) and c) with the DF module.

TABLE IV
AVERAGE TRAINING TIME FOR DIFFERENT METHODS

| | 5-shot-2C | 5-shot-5C | 10-shot-2C | 10-shot-5C |
|---|---|---|---|---|
| MN [29] | 125.23 (s) | 652.65 (s) | 432.25 (s) | 3025.65 (s) |
| PN [30] | 118.16 (s) | 752.45 (s) | 354.72 (s) | 2546.36 (s) |
| RN [31] | 156.24 (s) | 819.65 (s) | 495.25 (s) | 3432.24 (s) |
| MAML [43] | 245.24 (s) | 792.68 (s) | 419.88 (s) | 2653.24 (s) |
| HetNet [67] | 126.54 (s) | **252.32 (s)** | 198.27 (s) | 1025.56 (s) |
| SFENet [68] | 87.26 (s) | 256.78 (s) | **175.36 (s)** | **986.71 (s)** |
| **EmoDSN** | **76.53 (s)** | 419.25 (s) | 269.54 (s) | 1543.25 (s) |

As shown in Fig. 9, after the correction of labels, the accuracies increase 19.12% on average of three datasets. Since both no-CL and CL use the simple averaging distance learned by the DSN, the detection of mislabeled samples can promote the classification performance of *EmoDSN*. However, we also find that using DF can result in an average acc increase of 8.43% compared with using CL. In addition, the performance of DF is more stable than CL: the average SD of DF is 5.26% lower than CL. The difference between the network with DF and CL is that DF uses a soft weighted average of D to estimate the emotion label. CL uses an arithmetic average of D after correcting the labels of the samples whose $P(s_{nm} \in l_m)$ <0.5. Thus, for few-shot learning based fine-grained emotion recognition, assigning low weights for an inexactly labeled sample can result in better and more stable performance compared with simply correcting it according to the intra data distribution of training samples (i.e., whether the distribution of this sample is coherent with others).

### I. Running Time and Efficiency

The average training time for different methods are shown in Table IV. Our model is implemented using Keras and Tensorflow. All our experiments are performed on a desktop with NVIDIA RTX 2080Ti GPU with 16 GB RAM. The one-to-many learning structure is used by MN, PN, RN and MAML. Thus, the number of training samples for them is $n(n-1)\cdot k^2$, where $n$ and $k$ are the numbers of shots and classes of the learning task respectively. Our method uses the pair-by-pair learning structure. The number of training samples is $\Sigma_{i=1}^{n\cdot k-1}(n\cdot k-i)$. For the fully-supervised learning methods, training samples are directly

input into the network without combining them into different pairs. Thus, the number of training samples for them is $n \cdot k$.

Although the fully-supervised methods have a more complex structure, the number of samples for training is less than FSL methods. Thus, their training time is shorter than FSL methods. However, they do not achieve up-to-chance level acc because their learning structures are not designed for converging on a small amount of training samples. For the FSL methods, the pair-by-pair learning structure used by our method results in fewer training samples compared with other FSL methods using one-to-many structures. Thus, our method requires less training time: *EmoDSN* requires only 54.83% of the average training time of other FSL methods. The result demonstrates the good efficiency of *EmoDSN* compared with baselines using both fully-supervised and FSL methods.

Although the result of 10-shot is better, it requires much more training time compared with 5-shot. As shown in Table IV, training the model using 10-shot takes almost 4 times as long as training the model using 5-shot. The testing acc and m-F1 score however, only increases 0.15 and 0.20% on average for the three datasets. Increasing the training samples from 1-shot to 5-shot however, result in the increase of acc and m-F1 for 11.58 and 9.32% respectively. Thus, using 5-shot makes a trade-off between training time and model accuracies.

## VI. Discussion

### A. Reaction Delay of Continuous Annotation

According to the research of Metallinou *et al.* [73], there are time delays (e.g., due to gender, age, distraction levels) between the occurrence of an emotional event and its annotation considering that continuous annotations are performed in real-time. If we use misaligned annotation as labels to train the network, it will overfit or not converge. Most of the previous works [23]–[25] use visual features from video stimuli to align the annotation. In these approaches (also known as *explicit compensation* [74]), the delay compensation and the emotion prediction are performed separately. However, these approaches assume that the reaction delay is fixed for different users watching the same video stimuli. This assumption is untenable as the reaction time is both stimulus dependent and individual dependent [74].

The last layer of *EmoDSN* can identify which sub-instances (signal segments with different time of delay) can better predict the fine-grained emotion labels. Once the network is trained, we can observe the instance gains in the last layer to find out with how much delay the network can perform the best. Our approach belongs to the *implicit compensation* [74], which compensates for delays while modeling the relationship between input signals and emotion labels. The uniqueness of our approach is that we do not have to manually adjust the parameters (e.g., the width of analysis window for LSTM [26] or the receptive field for CNN [75]) in the network for compensating different delays for different individuals.

To obtain the range of reaction delay, we first run the 1D-2 C task and get the delays of the sub-instances with maximum instance gain (i.e., have the highest probability to predict emotion labels). We follow the procedure of previous works [23], [25] that estimate the delay of each dimension (valence and arousal) separately. Fig. 10 shows the box plot of reaction delays estimate by *EmoDSN* for three datasets.

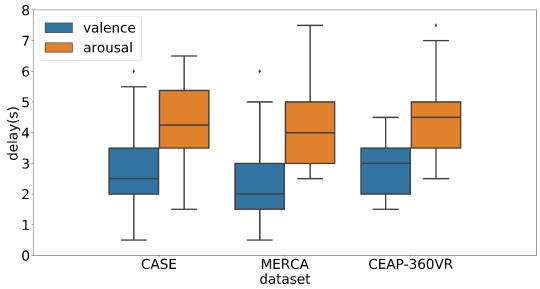


Fig. 10. The reaction delays for valence and arousal respectively.

The mean and standard deviation of delays are: CASE = 2.59(1.47), MERCA = 2.50(1.43), CEAP-360VR = 2.89(1.03) and CASE = 4.05(1.45), MERCA = 4.21(1.42), CEAP-360VR = 4.38(1.27) for valence and arousal respectively. The mean delay for arousal is higher than the delay for valence for all the three datasets. A Shapiro-Wilk test shows that the delays for both valence and arousal in three datasets are all normally distributed (all $p > 0.05$ for three datasets). For the comparison between different scenarios (desktop, mobile and VR for CASE, MERCA and CEAP-360VR respectively), we perform a ANOVA. Here we do not find a significant effect of scenarios on both valence ($F(2, 80) = 0.795, p = 0.455, \eta_p^2 = 0.019$ and arousal ($F(2, 80) = 0.416, p = 0.661, \eta_p^2 = 0.010$. However, through Welch's t tests, we do find there is significant difference between the delay of valence and arousal for CASE (t(58)=2.869, p <0.01, Cohen's $d = 0.944$), MERCA (t(40)=3.804, p <0.01, Cohen's $d = 1.372$) and CEAP-360VR (t(62)=5.045, p <0.01, Cohen's $d = 1.340$) respectively.

These results show that users need more time to react for annotating arousal than valence. This finding is coherent with most of the previous works using *explicit* [23]–[25] compensation methods. The averaged delays (2.66 s and 4.21 s for V-A) obtained by our method are also similar to the results obtained by *explicit* methods (e.g., 2 s and 4 s from Huang *et al.* [23], 3.08 s and 3.95 s from Marioryad *et al.* [24] for V-A respectively). Thus, our method for compensating reaction delay can provide similar results without using visual and audio features from stimuli. The average annotation delays in different datasets collected in different scenarios are comparable. The reason for this finding is that the annotations of all three datasets were collected using the joystick-based annotation interface.

We also conduct an experiment to find out whether sliding windows with long delays can introduce redundant information from other temporal moments for emotion recognition. Fig. 11 shows the relationship between the steps of sliding windows and recognition acc for 1D-2 C arousal and valence respectively. The recognition acc keeps increasing for both valence and arousal recognition when the steps of delay increase from 0 s to 7 s. The low acc caused by the short delay time of sliding windows show that if the sliding windows cannot cover enough delay, the embedding network will fail to identify the sub-instances

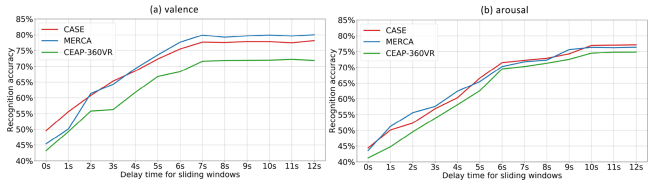


Fig. 11. The relationship between the steps of sliding windows and recognition acc.

which represent the emotion label of that moment. The recognition acc for V-A becomes stable after increasing the delay of sliding windows for 7 s and 10 s respectively. Thus, the noise of adding steps of sliding windows can be filtered by the MIL module: the recognition acc do not decrease when we add the steps of the sliding window. Instead of fully-supervised learning, all the sub-instances are weakly supervised by the emotion labels. The weights learned by MIL layers represent the probability of one sub-instance for discriminating samples between different emotion categories. Thus, the redundant information from other temporal moments can be automatically filtered (i.e., assign low weights). The results are also in line with the finding that the annotation delay of arousal is higher than the delay of valence: we need to add more steps (i.e., delay time) of sliding windows to cover the corresponding sub-instances for arousal recognition.

### B. Do the Temporal Moments of Training Samples Affect the Performance?

In Section V, we randomly sample $N$ training samples from each emotion category to train *EmoDSN*. Although it is the standard evaluation procedure to test few-shot learning algorithms, it is difficult to get randomly balanced number of samples with different emotion labels. When applying the algorithm for evaluating the user experience of watching videos, the possible methods are 1) randomly stop the video and ask users to annotate their emotions or 2) ask users to annotate at some fixed temporal moments to obtain the emotion labels for training. Since we use only few annotated samples to train the network, we want to find out samples from which temporal moments can better represent the distribution for the whole video watching and result in better recognition results. We also want to explore the amount of samples *EmoDSN* needs to obtain accurate recognition when selecting training samples in different temporal moments of video watching. Answering these two questions can help researchers maximize the performance of *EmoDSN* and minimize the amount of training samples by asking users to annotate at the most suitable temporal moments in video watching.

To achieve this, we select training samples from both fixed and random temporal moments of video watching and compare the recognition acc (1D-2 C) when training with different amounts of samples. Specifically, we choose the beginning, ending and the changing points as fixed temporal moments and compare the result with the random moments:

- *Beginning:* We choose the first $K$ samples from a video watching as training samples and test on the rest.
- *Ending:* We choose the last $K$ samples from a video watching as training samples and test on the rest.
- *Changing points:* According to the research of Sharma *et al.* [32], the changing points in continuous annotation can signify emotionally salient moments. Thus, we want to find out whether these samples can better represent the distribution for the whole video watching. We select samples from the changing points of annotation (obtained using the Changing Points Analysis (CPA) [32]) as the training samples and test on the remaining samples.
- *Random:* We randomly choose $K$ samples from one video watching as the training samples and test on the remaining samples. Unlike balanced random selection in Section V, it does not ensure each emotion category has a balanced number of training samples.

The results of how the acc of *EmoDSN* changes with different amounts of samples from different temporal moments of video watching is shown in Fig. 12. From Fig. 12(d) we observe that random selection results in great fluctuation of the recognition acc when more samples are used for training. Selecting from fixed temporal moments however (Fig. 12(a)–(c)), results in relatively stable performance when inputting more training samples. Thus, selecting training samples from fixed temporal moments can result in more stable performance when we only use few annotated samples for training.

We also observe that if we choose training samples from the beginning of video watching, the algorithm needs more training samples to converge. It needs more than 10 training samples (20 seconds) to increase the recognition acc above 50%. Using the ending moments however, requires less than 8 training samples (16 seconds) to achieve 50% acc. The best temporal moments to select the training samples are the changing points: the acc exceeds 70% by only using 4 samples (8 seconds) for training. Thus, the samples at the changing points and the ending moments can better represent the distribution of the whole video watching and result in better acc with fewer samples.

The results we obtain are coherent with the *peak-end theory* [76] that the most salient (peak) or recent (end) moments can better represent the emotions of users while watching videos. We also observe that the distributions of samples with specific emotion labels are different across the temporal moments. Fig. 13 shows the percentage of samples with high/low V-A labels in different temporal moments of video watching. Compared with the ending moments of video watching, most of the samples (more than 70% for all three datasets) from the beginning moments are labeled as high V-A. If we choose these samples as training data, the imbalanced training set can result in mis-convergence of the learning network (e.g., in Fig. 12(a) when the amount of samples $<$12 s, acc $<$30%). It also explains why fewer training samples are required from the end of video watching for good results: the samples are more balanced at the end of video watching.

In conclusion, the temporal moments of training samples do have influence on the performance of *EmoDSN*. The take-way message from this experiment is that samples from the changing points of emotion and the ending moments of video watching are better training samples when only few samples are available for building up an emotion recognition system.

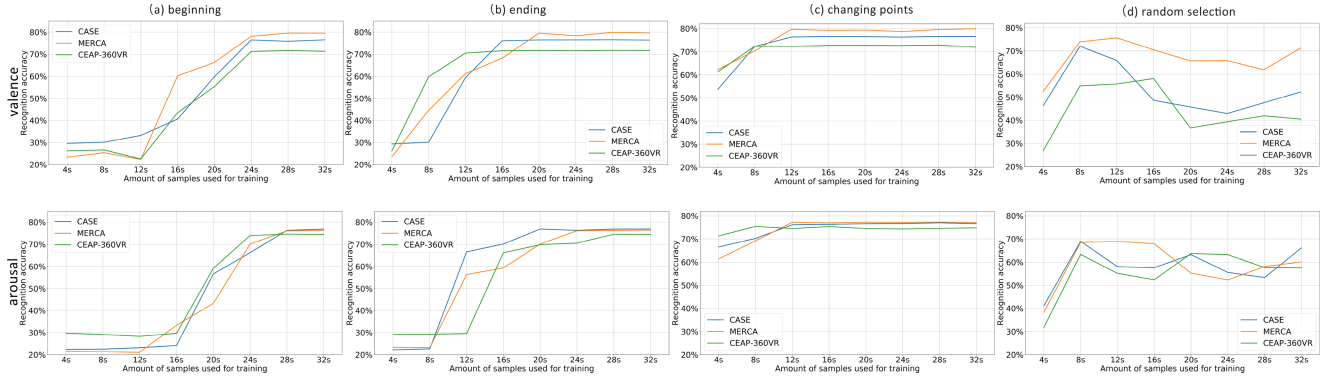


Fig. 12. The 1D-2 C recognition acc when training *EmoDSN* with samples from the (a) beginning, (b) ending, (c) changing points and (d) random position of video watching. The granularity of samples is 2 seconds. The amount of samples are shown in the unite of seconds (e.g., 4 seconds = 2 samples).

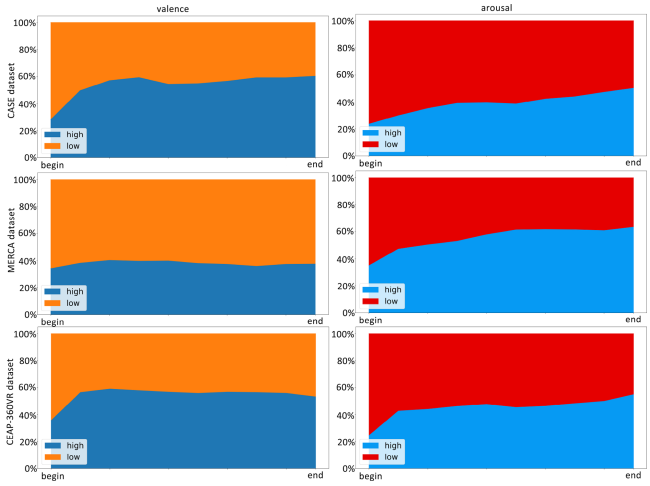


Fig. 13. Percentage of samples with high/low V-A labels in different temporal moments of video watching for CASE, MERCA and CEAP-360VR.

## VII. Limitations and Future Work

Given the challenges of predicting valence and arousal labels at a fine level of granularity using only few annotated samples, there are natural limitations to our work. First, *EmoDSN* only works well for the personalized or subject-dependent emotion recognition model. Since the patterns of physiological signals are highly variable between subjects [39], [63], using few annotated samples to model it is challenging and relies on the careful selection of training samples. In the future, we will extend *EmoDSN* for subject-independent emotion recognition model by finding out which training samples can represent the inter-subject variability of physiological signals. In addition, *EmoDSN* requires discretization of continuous labels for fine-grained recognition since *EmoDSN* is designed specifically for classification instead of regression. In the future, we will extend the *EmoDSN* into few-shot regression [77] algorithm and obtain continuous output for emotion recognition. It is also essential for us to compare the performance of *EmoDSN* on more datasets to further test its generalizability. However, the number of datasets with continuously annotated physiological signals is to date limited. It lacks benchmark results using basic few-shot learning methods. Thus, it is difficult to make comparisons with more few-shot learning methods.

## VIII. Conclusion

Fine-grained emotion recognition requires training the algorithm with large amounts of continuous emotion labels. In this paper, we propose *EmoDSN*, a Deep Siamese Network based few-shot learning algorithm to classify fine-grained valence and arousal with only a small amount of annotated signals. The embedding network of *EmoDSN* enables our algorithm to compensate the reaction delay of annotation while predicting the fine-grained valence and arousal. The distance fusion module of *EmoDSN* minimizes the overfitting problem caused by mislabeled training samples. The proposed algorithm achieves reasonable performance (averaged accuracy of 76.04, 76.62 and 57.62% for 1D-2 C valence, 1D-2 C arousal and 2D-5 C respectively) by using only 5 shot as training data for subject-dependent testing on three datasets collected in three different environments (i.e., desktop, mobile, and HMD-based VR). Our algorithm also outperforms other few-shot learning algorithms which are widely used for emotion recognition. The ablation study shows that the embedding network and distance fusion module, which are specifically designed for physiological signals based fine-grained emotion recognition, can significantly improve the recognition accuracy. Our experiment on reaction delay of annotation shows that 1) the reaction delay for arousal is longer than the delay for valence and 2) the reaction delays between different scenarios have no significant difference. We also find that the changing points of emotion annotation and the ending moments of video watching are better temporal moments for selecting training samples: if we select training samples from these two temporal moments, *EmoDSN* can provide better recognition results with fewer annotated samples. Source code for EmoDSN is publicly available on https://github.com/cwi-dis/EmoDSN.

## References

[1] 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.

[2] 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 Interact.*, 2019, pp. 404–408.

[3] M. A. Jarwar and I. Chong, "Web objects based contextual data quality assessment model for semantic data application," *Appl. Sci.*, vol. 10, no. 6, 2020, Art. no. 2181.

[4] 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.

[5] E. Paul, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York, NY, USA: OWL Books, 2007.

[6] R. W. Levenson, "Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity," *Social Psychophysiology: Theory Clin. Appl.*, Hoboken, NJ, USA: Wiley, 1988, pp. 17–42.

[7] 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.

[8] 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.

[9] 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, Jan.–Mar. 2016.

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

[11] 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.

[12] W. James, "What is an emotion?," *Mind*, vol. 9, no. 34, pp. 188–205, 1884.

[13] 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.

[14] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.

[15] 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.

[16] 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.

[17] 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.

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

[19] C. Y. Park *et al.*, "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Sci. Data*, vol. 7, no. 1, 2020, Art. no. 293.

[20] Y. Wang and Q. Yao, "Few-shot learning: A survey," 2019, *arXiv:1904.05046*.

[21] S. Jiang, F. Firouzi, K. Chakrabarty, and E. Elbogen, "A resilient and hierarchical IoT-based solution for stress monitoring in everyday settings," *IEEE Internet Things J.*, to be published, doi: 10.1109/JIOT.2021.3122015.

[22] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, "Zero-shot emotion recognition via affective structural embedding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1151–1160.

[23] Z. Huang *et al.*, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge*, 2015, pp. 41–48.

[24] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Trans. Affect. Comput.*, vol. 6, no. 2, pp. 97–108, Apr.–Jun. 2015.

[25] 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. 1–5.

[26] D. Le, Z. Aldeneh, and E. M. Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network," in *Proc. Interspeech*, 2017, pp. 1108–1112.

[27] A. Patane and M. Kwiatkowska, "Calibrating the classifier: Siamese neural network architecture for end-to-end arousal recognition from ECG," in *Proc. Int. Conf. Mach. Learn., Optim., Data Sci.*, 2018, pp. 1–13.

[28] 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.

[29] O. Vinyals *et al.*, "Matching networks for one shot learning," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29, pp. 3630–3638.

[30] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," 2017, *arXiv:1703.05175*.

[31] F. Sung *et al.*, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

[32] 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. Affect. Comput.*, vol. 11, no. 1, pp. 78–84, Jan.–Mar. 2020.

[33] G. Chen, Y. Zhu, Z. Hong, and Z. Yang, "EmotionalGAN: Generating ECG to enhance emotion state classification," in *Proc. Int. Conf. Artif. Intell. Comput. Sci.*, 2019, pp. 309–313.

[34] Q. Zhong, Y. Zhu, D. Cai, L. Xiao, and H. Zhang, "Electroencephalogram access for emotion recognition based on a deep hybrid network," *Front. Hum. Neurosci.*, vol. 14, 2020, Art. no. 589001.

[35] Y. Jiao, Y. Deng, Y. Luo, and B.-L. Lu, "Driver sleepiness detection from EEG and EOG signals using GAN and LSTM networks," *Neurocomputing*, vol. 408, pp. 100–111, 2020.

[36] R. Song *et al.*, "PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1373–1384, May 2021.

[37] 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.

[38] P. Salehi, A. Chalechale, and M. Taghizadeh, "Generative adversarial networks (GANs): An overview of theoretical model, evaluation metrics, and recent developments," 2020, *arXiv:2005.13178*.

[39] 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, Jan.–Mar. 2022.

[40] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, "Weakly-supervised learning for fine-grained emotion recognition using physiological signals," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2022.3158234.

[41] E. Pei, D. Jiang, M. Alioscha-Perez, and H. Sahli, "Continuous affect recognition with weakly supervised learning," *Multimedia Tools Appl.*, vol. 78, no. 14, pp. 19387–19412, 2019.

[42] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge*, 2015, pp. 65–72.

[43] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[44] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, 2018, pp. 400–408.

[45] J. Bromley *et al.*, "Signature verification using a 'Siamese' time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, 1993.

[46] W. Hayale, P. S. Negi, and M. Mahoor, "Deep siamese neural networks for facial expression recognition in the wild," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2021.3077248.

[47] A. Mehmood, M. Maqsood, M. Bashir, and Y. Shuyuan, "A deep siamese convolution neural network for multi-class classification of alzheimer disease," *Brain Sci.*, vol. 10, no. 2, 2020, Art. no. 84.

[48] K. Feng and T. Chaspari, "A siamese neural network with modified distance loss for transfer learning in speech emotion recognition," 2020, *arXiv:2006.03001*.

[49] 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. Interact.*, 2013, pp. 73–78.

[50] 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.

[51] 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.
[52] 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.
[53] A. V. D. Oord *et al.*, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
[54] 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.
[55] 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.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[57] 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.
[58] 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.
[59] R. W. Daniels, *Approximation Methods for Electronic Filter Design: With Applications to Passive, Active, and Digital Networks*. New York, NY, USA: McGraw-Hill, 1974.
[60] 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.
[61] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on deap dataset," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, 4746–4752.
[62] Y. Li, J. Huang, H. Zhou, and N. Zhong, "Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks," *Appl. Sci.*, vol. 7, no. 10, 2017, Art. no. 1060.
[63] 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.
[64] J. Na, H. Jung, H. J. Chang, and W. Hwang, "FixBi: Bridging domain spaces for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1094–1103.
[65] Y. Wang, S. Qiu, X. Ma, and H. He, "A prototype-based SPD matrix network for domain adaptation EEG emotion recognition," *Pattern Recognit.*, vol. 110, 2021, Art. no. 107626.
[66] M. Chen *et al.*, "Diversity transfer network for few-shot learning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 10559–10566.
[67] Z. Jia *et al.*, "HetEmotionNet: Two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1047–1056.
[68] X. Deng, J. Zhu, and S. Yang, "SFE-Net: EEG-based emotion recognition with symmetrical spatial feature extraction," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 2391–2400.
[69] H. Zhang, N. M. Nasrabadi, T. S. Huang, and Y. Zhang, "Transient acoustic signal classification using joint sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 2220–2223.
[70] S. Golkar, M. Kagan, and K. Cho, "Continual learning via neural pruning," 2019, *arXiv:1903.04476*.
[71] P. Kumar *et al.*, "A statistical significance of differences in classification accuracy of crop types using different classification algorithms," *Geocarto Int.*, vol. 32, no. 2, pp. 206–224, 2017.
[72] M. Goldblum, L. Fowl, and T. Goldstein, "Robust few-shot learning with adversarially queried meta-learners," 2019, *arXiv:1910.00982*.
[73] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.
[74] 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.
[75] S. Khorram, Z. Aldeneh, D. Dimitriadis, M. McInnis, and E. M. Provost, "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition," 2017, *arXiv:1708.07050*.
[76] B. L. Fredrickson and D. Kahneman, "Duration neglect in retrospective evaluations of affective episodes," *J. Pers. Social Psychol.*, vol. 65, no. 1, pp. 45–55, 1993.
[77] Y. Loo, S. K. Lim, G. Roig, and N.-M. Cheung, "Few-shot regression via learned basis functions," 2019.

**Tianyi Zhang** (Student Member, IEEE) is currently working toward the Ph.D. degree with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands. He is associated with the Distributed & Interactive Systems Group, Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands, the national research institute for mathematics and computer science. His research interests include human-computer interaction and machine learning based affective computing.

**Abdallah El Ali** (Member, IEEE) received the Ph.D. degree from the University of Amsterdam, Amsterdam, The Netherlands, in 2013. He is currently a Research Scientist (tenured) with Distributed and Interactive Systems Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands. He is leading human-computer interaction research with a focus on affective interactive systems. His research interests include ground truth label acquisition techniques, affective state visualization across environments (mobile, wearable, and XR), and bio-responsive interactive prototypes.

**Alan Hanjalic** (Fellow, IEEE) is currently a Professor of computer science with the Delft University of Technology, Delft, The Netherlands, and the Founder and Head of the Multimedia Computing Group. He has authored or coauthored more than 150 publications in his research areas, which include multimedia information retrieval and recommender systems. He was a co-recipient of the Best Paper Award at the ACM Conference on Recommender Systems (RecSys) 2012, Multimedia Grand Challenge Award at the ACM International Conference on Multimedia (ACM Multimedia) 2015, and Best Paper Award at the ACM Conference on Multimedia, 2017. He is the Chair of the Steering Committee of the IEEE TRANSACTIONS ON MULTIMEDIA, an Associate Editor-in-Chief of the *IEEE Multimedia Magazine*, and a Member of the Editorial Board of the *ACM Transactions in Multimedia*, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, and *International Journal of Multimedia Information Retrieval*. He is also the General and Program Chair or Co-Chair of the organizing committees of leading conferences in the multimedia domain, including ACM Multimedia, ACM CIVR/ICMR, and IEEE ICME.

**Pablo Cesar** (Senior Member, IEEE) currently leads the Distributed and Interactive Systems Group, Centrum Wiskunde & Infomartica, Amsterdam, The Netherlands, and also a Full Professor with the Delft University of Technology, Delft The Netherlands. His research interests include human-computer interaction and multimedia systems, and focuses on modeling and controlling complex collections of media objects, including real-time media and sensor data that are distributed in time and space. He was the recipient of 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 the IEEE MULTIMEDIA, *ACM Transactions on Multimedia*, and IEEE TRANSACTIONS OF MULTIMEDIA. He was an Invited Expert with the European Commission's Future Media Internet Architecture Think Tank. graphical user interfaces.