Pairwise Ranking Network for Affect Recognition

Georgios Zoumpourlis
School of Electronic Engineering and Computer Science
Queen Mary University of London
g.zoumpourlis@qmul.ac.uk

Ioannis Patras
School of Electronic Engineering and Computer Science
Queen Mary University of London
i.patras@qmul.ac.uk

***Abstract*—In this work we study the problem of emotion recognition under the prism of preference learning. Affective datasets are typically annotated by assigning a single absolute label, i.e. a numerical value that describes the intensity of an emotional attribute, to each sample. Then, the majority of existing works on affect recognition employ sample-wise classification/regression methods to predict affective states, using those annotations. We take a different approach and use a deep network architecture that performs joint training on the tasks of classification/regression of samples and ordinal ranking between pairs of samples. By treating input samples in a pairwise manner, we leverage the auxiliary task of inferring the ordinal relation between their corresponding affective states. Incorporating the ranking objective allows capturing the inherently ordinal structure of emotions and learning the inter-sample relations, resulting in better generalization. Our method is incorporated into existing affect recognition architectures and evaluated on datasets of electroencephalograms (EEG) and images. We show that the approach proposed in this work leads to consistent performance gains when incorporated in classification/regression networks.**

***Index Terms*—Affect annotation, emotion recognition, electroencephalogram, facial expressions**

I. Introduction

Human emotional experiences have a critical role in our everyday lives. Reliable affect estimation is one of the main goals of the affective computing field, requiring multi-disciplinary research that spans across computer science, neuroscience and psychology. Affect can be analyzed by studying both the physical and neurophysiological changes that occur during emotion elicitation through facial expressions [7], [12], body gestures [21], speech [3], brain activity [29], etc. Various theories have been proposed to model emotions [25], [27], with the most common ones being the categorical emotion model of Ekman *et al*. [6] and the dimensional model of Russell [26]. Affective datasets adopt such theoretical models to derive their data annotations, either from experiment participants reporting self-assessment emotion ratings, or from offline external annotators.

During emotion data labelling, typically, humans assign a value in a continuous range, for each emotional behavior. These values are assumed to be on an absolute scale, however even for a single annotator the perception of the rating scale may change across time [16], while there are different subjective biases across multiple annotators [19]. Works inspired from the adaptation level theory of Helson [9], suggest that human judgments of presented stimuli are relative to the context [28], including previously encountered stimuli, rather than absolute. Therefore emotions can be expressed in relative terms, i.e. through comparisons between different affective state levels. Labelling emotions by assigning relative values has been an alternative path to the traditional scheme of absolute labels [14], [32]. This means that annotating emotions involves comparison of the human affective states between past and forthcoming experiences. Therefore, one possible way of inferring such ordinal relations between affective states, is through machine learning models that can explicitly compare them.

In this work, we study the problem of affect recognition on datasets where annotations are provided in the form of sample-wise labels. Typically, plain regression or classification approaches are applied on such datasets. In the case of regression, the inherent biases of continuous affect annotations described above, are harmful for the training process thus also for the model's performance [34]. Other problems arise when adopting classification approaches as a remedy to the shortcomings of regression. Discrete classes cannot express the compoundness of emotions. Transforming ratings of ordinal nature into nominal classes results in information loss regarding the structure of ratings. Furthermore, the class splitting criteria defined by researchers, do not always accurately reflect the manifestations of affect [16]. Hence, a more suitable approach is preference learning [8], that involves comparing emotions. The superiority of preference learning methods over classification algorithms for affect recognition, has been previously studied in [18]. We follow an alternative direction, investigating the utilization of preference learning as an auxiliary objective to improve the performance of deep neural networks on classification/regression.

Despite the exciting results of deep learning methods on affective computing problems, the possibility of building deep networks that can compare samples corresponding to different affective states, has remained mostly unexplored. Refraining from using solely a sample-wise classification/regression objective, we propose employing an additional pairwise objective, namely the emotional rating comparison. Considering a pair of data samples and their affective labels, the comparison task infers the ordinal ranking relation between the labels of the samples (i.e. higher/similar/lower arousal,

The work of Georgios Zoumpourlis was supported by QMUL Principal's Studentship. This work was also supported by EU H2020 project AI4Media No. 951911. We gratefully acknowledge NVIDIA for the donation of the GTX Titan X GPU used for this research.

978-1-6654-0019-0/21/$31.00 ©2021 IEEE

higher/similar/lower valence). We use a shared deep feature extractor along with separate network heads that infer the affective state level of each sample and perform pairwise ranking between samples. Our experiments show that the former task benefits from the latter, as treating the data in a pairwise manner enables better representation learning. The main contributions of our work are the following:

- We propose a deep architecture that is jointly trained on sample-wise classification/regression and pairwise ordinal ranking.
- We conduct experiments on neurophysiological and visual data, showing consistent gains from the incorporation of a ranking objective in the training process.
- We perform ablation studies on models trained with our proposed method, to quantitatively evaluate the benefits of various components.

II. Related work

In this Section, we present an overview of the related work on the topics of affect modelling and affect recognition. These two areas are closely connected, therefore meaningful combinations of knowledge from both of them can lead to new insights.

A. Affect modelling

Affect modelling [4] is the task of mapping inputs (e.g. extracted features from facial expressions or neurophysiological signals) to an affective state. The kind of affect annotation determines the expected output of the model, hence also the type of machine learning approach that can be applied, namely regression, classification or preference learning [8], [24]. The subjectivity of emotional rating annotations can be observed across multiple experimental sessions of an individual, or across multiple human participants, making the annotation process inherently biased. Treating ratings as numerical values through regression, leads to trained models of questionnable performance. Converting annotations into discrete classes (i.e., splitting the scale range of ratings into classes such as "low"/"high", by defining a threshold value), results in models that ignore the ordinal relation, not only across different classes, but also between samples within each class.

Emotion recognition methods could benefit by favoring relative over absolute schemes, for representing and annotating emotions. Research findings that highlight the reliability of collecting rank-based affective annotations, pave the way for utilising such datasets for emotion analysis. Yannakakis *et al*. [33] proposed a discrete rank-based real-time affect annotation tool, showing higher quality of the obtained relative labels compared to absolute labels. In [17], annotations collected using a relative and unbounded labelling scheme yielded high inter-rater agreement. Differently from these works, we consider existing datasets having annotations that are *not* rank-based, and suggest converting their original absolute labels into pairwise ranking labels.

B. Affect recognition

Recently, a broad range of methods have been proposed to address emotion recognition tasks on several modalities. In our work we focus on electroencephalograms (EEG) and facial expression images as inputs, experimenting on widely used datasets. Commonly, attempts of training stronger facial expression recognition models are based on aspects irrelevant to annotations per se. Panda *et al*. propose training on large-scale data crawled from the internet [22], while Kollias *et al*. artificially generate facial images [11] to be included in the training set. Studying the temporal dynamics of emotions [1] or applying state-of-the-art deep learning architectures such as Transformers [15] is another line of research. The study of Zhang *et al*. [35] investigates the problem of noisy labels on affective datasets, indicating the significance of dealing properly with such annotations. In EEG-based affective experiments, the annotations are typically obtained through self-assessment ratings of the human participants. Often, evaluation protocols impose quantizing these ratings into classes corresponding to different affective state levels [10]. In [13], the subject-specific range of each individual's ratings is studied, and the classes are reformulated by computing a personalized threshold. Overall, the development of affect recognition models that abide to the nature of emotions is of paramount importance to the affective computing field.

III. Proposed method

The main motivation of our work is to investigate meaningful combinations of classification/regression and ordinal ranking through deep neural networks, in the field of affective computing. In contrast to typical network architectures that operate on affective data solely in a sample-wise manner, we aim to additionaly perform pairwise operations between samples, learning the ordinal relation between their corresponding affective ratings. Our goal is to boost the performance of emotion recognition models on their classification/regression measures, leveraging the additional supervision of a ranking task only during training. Traditional preference learning systems such as RankNet [2], train a function that maintains a higher score for the preferred option. The preference decision is a fixed operation on sample-wise preference scores, without involving any trainable parameter. Our method differs from such systems, as it learns the ordinal relation through a trainable module. Considering that the emotion label space has an inherently ordinal structure, we avoid disregarding such knowledge, by further exploiting it through the ranking task. To achieve this, we utilise the provided affective state annotations to form rank-based labels, and construct a deep architecture that can handle both the end-goal task of classification/regression, as well as the additional task of pairwise ranking. In the following paragraphs, we explain various aspects of our method.

A. Methodology

Pairwise ordinal ranking: The proposed methodology that derives pairwise ranking labels is applicable on datasets having

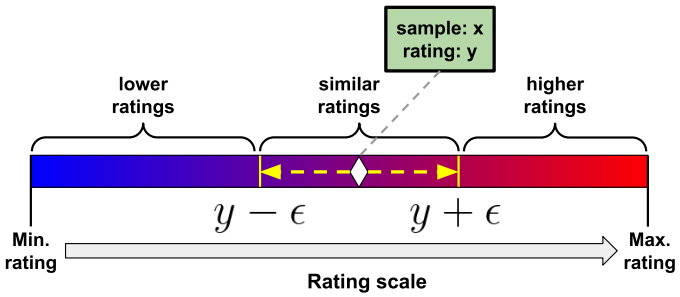

Fig. 1: Illustration of the ordinal relations defined over a bounded continuous rating scale.

as annotations either *continuous* affective ratings or *categorical* labels of ordinal nature. We explain its functionality on continuous labels, defined on a bounded scale. Considering a pair of samples x_1 and x_2 (with corresponding affective rating labels y_1 and y_2), the goal of the ranking task is to infer the ordinal relation between the labels y_1 and y_2. In previous works this is addressed by establishing a preference of the sample with the higher rating over the other sample, i.e. $x_1 \succ x_2$ or $x_1 \prec x_2$. The symbols of "$\prec$"/"$\succ$" denote preceding/succeeding order of the samples with respect to their ratings y_1 and y_2, i.e. by using these symbols we do not imply a comparison on the raw feature values of x_1 and x_2. A minimum difference value between the compared ratings is used to discard unclear comparisons. To avoid posing very strict constraints over pairs of ratings with small difference, we opt to add a third case of rank, namely the case $x_1 \sim x_2$, if x_1 and x_2 have similar ratings [23]. We define a hyperparameter $\epsilon > 0$, called "rank tolerance", such that $x_1 \sim x_2$ holds true when $|y_1 - y_2| \leq \epsilon$. Thus, $x_1 \succ x_2$ when x_1 has a higher rating than x_2 under the condition $y_1 > (y_2 + \epsilon)$, and $x_1 \prec x_2$ when $y_1 < (y_2 - \epsilon)$. The ordinal relations for continuous ratings are shown in Table I, as well as in Fig. 1.

Relation	Condition
$x_1 \succ x_2$	$y_1 > (y_2 + \epsilon)$
$x_1 \sim x_2$	$\|y_1 - y_2\| \leq \epsilon$
$x_1 \prec x_2$	$y_1 < (y_2 - \epsilon)$

TABLE I: List of ordinal ranking relations and their corresponding conditions, when performing a comparison operation over continuous ratings.

Joint training - combining ranking with end-goal tasks: Our method is simple and it can be integrated into existing affect recognition architectures. In essence, every deep neural network operating on an end-goal task of affect classification/regression, consists of a backbone that extracts feature representations which are ultimately fed into a classification/regression layer. We suggest adding an extra supervisory signal, by imposing a pairwise ranking objective on the intermediate representations learned by the backbone, leveraging the knowledge around the ordinal nature of emotions. The ranking task is performed by a ranking head that is stacked on top of the backbone network. The processing pipeline for classification/regression remains intact and the total architecture is trained in an end-to-end manner. We fully backpropagate the gradients of both the classification/regression loss and ranking loss to the backbone, updating its weights based on both loss terms. The backbone network benefits from the additional ranking supervision, extracting features that enable better generalisation on the end-goal task. The classification and ranking losses are computed using a cross-entropy criterion, while the regression loss is computed using a Mean Squared Error (MSE) criterion.

B. Network architecture

We aim to build an architecture that operates on affective data inputs to perform sample-wise classification/regression of emotions, as well as a pairwise comparison operation (ordinal ranking) with respect to the emotional ratings for a pair of samples. Regarding the implementation of deep networks that accomodate pairwise operations, our work builds on the Relation Networks [30] that have been used for few-shot image recognition. In the context of Relation Networks, a *relation* module refers to a mechanism that learns to compare feature embeddings for a pair of samples, to determine whether they have the same class label or not. We adapt the framework of [30] to suit the purposes of pairwise ranking. We propose using a *ranking* module that learns to perform ordinal ranking on the feature embeddings of a pair of samples, by inferring the ordinal relation between their affective ratings. Note that the inputs of our ranking module are pairwise feature embeddings, formed by concatenating the sample-wise embeddings obtained from a backbone feature extractor, for each pair of samples. Our architecture, named Pairwise Ranking Network ("PRNet"), can be seen in Fig. 2. A detailed explanation of its consisting modules is provided below.

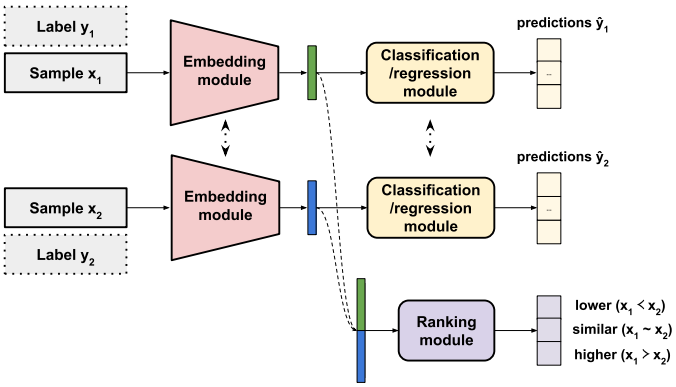

Fig. 2: The architecture of a Pairwise Ranking Network that accomodates joint training on classification/regression and ranking tasks.

Embedding module: The embedding module is the backbone of our architecture, serving as a feature extractor. The batch samples are fed as inputs to the embedding module and a feature embedding is computed for each sample. The produced embeddings are to be further processed for the tasks

of classification/regression and ranking, by the corresponding modules.

Classification/regression module: The classification/regression module receives as input the features produced by the embedding module, and predicts the affective state for each sample. In the classification scenario, the groundtruth targets are discrete emotion classes (e.g. "low"/"high" arousal, "low"/"high" valence) while in the regression scenario, the targets are the original arousal/valence annotations in a continuous space. We denote the classification and regression predictions as $\hat{\mathbf{y}}_{\text{cls}}$ and $\hat{\mathbf{y}}_{\text{regr}}$ respectively. Similarly, the corresponding groundtruth values are $\mathbf{y}_{\text{cls}}$ and $\mathbf{y}_{\text{regr}}$. Note that $\hat{\mathbf{y}}_{\text{cls}}$ contains probabilities obtained by passing the outputs of the classification module through a softmax layer, while $\mathbf{y}_{\text{cls}}$ contains one-hot encodings of the labels. The loss terms $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{regr}}$ of the classification and regression tasks are defined as follows:

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{k_{\text{cls}}} \mathbf{y}_{\text{cls}}^{\text{i}} \log(\hat{\mathbf{y}}_{\text{cls}}^{\text{i}}) \tag{1}$$

$$\mathcal{L}_{\text{regr}} = \frac{1}{k_{\text{regr}}} \sum_{i=1}^{k_{\text{regr}}} (\mathbf{y}_{\text{regr}}^{\text{i}} - \hat{\mathbf{y}}_{\text{regr}}^{\text{i}})^2 \tag{2}$$

Ranking module: The ranking module operates on pairwise feature representations that correspond to sample pairs, and infers their ordinal relation with respect to their affective ratings. To form the pairwise feature representation of two samples, we get the feature vectors extracted from the embedding module for both samples, and we concatenate them across the channel dimension. To form multiple pairs of sample embeddings during training with a batch size of N_{b}, we split each batch into two sub-batches of size $N_{\text{sub}} = \frac{N_{\text{b}}}{2}$. Every sample of each sub-batch is compared against all samples of the other sub-batch, yielding $(N_{\text{sub}})^2$ pairs in total. Denoting the softmaxed ranking predictions and one-hot groundtruth values as $\hat{\mathbf{y}}_{\text{rank}}$ and $\mathbf{y}_{\text{rank}}$ respectively, the loss term $\mathcal{L}_{\text{rank}}$ of the ranking task is defined as follows:

$$\mathcal{L}_{\text{rank}} = -\sum_{i=1}^{k_{\text{rank}}} \mathbf{y}_{\text{rank}}^{\text{i}} \log(\hat{\mathbf{y}}_{\text{rank}}^{\text{i}}) \tag{3}$$

The total loss that is used to optimize the Pairwise Ranking Network is the sum of the loss on the end-goal task and the ranking loss. We use a coefficient α to weight the contribution of the ranking loss to the total loss, i.e. $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha\mathcal{L}_{\text{rank}}$ in the case of classification or $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{regr}} + \alpha\mathcal{L}_{\text{rank}}$ in the case of regression. When not stated otherwise, we set the value of α equal to 1.

Architecture details: For our experiments on EEG data, the embedding module consists of two fully-connected (FC) layers with 128 nodes each, receiving 100-dimensional feature vectors as inputs. The classification module consists of one FC layer for each of the targets (i.e, Arousal, Valence), with k_{cls} output nodes, where k_{cls} is the number of classes. The ranking module consists of one FC layer for each of the targets, having $k_{\text{rank}} = 3$ output nodes.

When operating on visual data, the embedding module consists of five convolutional stages, each stage having two convolutional layers and a max-pooling layer with a downsample rate of 2. The number of channels for the convolutional stages is $\{64, 128, 256, 512, 512\}$, and the feature maps of the last stage are flattened, so that the computed embeddings are in the form of feature vectors. The regression module consists of two FC layers, having 256 and k_{regr} output nodes, where $k_{\text{regr}} = 2$ is the number of regression targets (arousal, valence). The ranking module has two FC layers for each of the targets, having 256 and $k_{\text{rank}} = 3$ output nodes.

The baseline model for our experiments is the composition of the embedding and classification/regression modules, i.e. a simple model with a feature extractor and a classifier/regressor.

IV. Experimental results

We apply our method on three emotion recognition problems where the original affective annotations are inherently ordinal, aiming to exploit this property through our analysis. Specifically, we study the datasets of DEAP [10], SEED [36] and AffectNet [20]. Each dataset has been annotated through a different process, and is evaluated on a different end-goal task. Investigating whether such tasks can benefit from pairwise ranking through a joint training, is an interesting direction of research. An overview of the datasets used in our study is shown in Table II.

A. Dataset details

DEAP dataset: DEAP [10] is a dataset for EEG-based emotion recognition, having 32 participants and 40 music video clips as stimuli, with a fixed duration of 60 seconds for each clip. Groundtruth labels for arousal and valence are given as self-assessment ratings in the continuous range of $[1.0, 9.0]$. The end-goal task on DEAP is the classification of "Low"/"High" affective states, defined by thresholding the rating scale in the middlepoint of 5.0. The classification head of our deep architecture predicts class scores for these two outputs on arousal and valence.

In the case of DEAP dataset, the original labels are continuous ratings and the end-goal task is classification. The affective ratings are quantized thus the ordinality of the initial labels is lost. Moreover, collapsing entire ranges of the rating scale into single classes leads to models that cannot reason about intra-class sample differences. The application of a ranking approach on the original ratings is straight-forward, following the ordinal relations that are shown in Table I.

SEED dataset: SEED [36] is a dataset for EEG-based emotion recognition, having 15 participants and 15 Chinese movie videos as stimuli, with varying duration for each clip (4 minutes in average). The labels are categorical, belonging in three classes, namely "Positive", "Neutral" and "Negative". The end-goal task of SEED is the classification of these three states.

Dataset	Modality	Annotation process	Annotation values	End-goal task
DEAP [10]	EEG	-Self-assessment reports -Varying per participant	Arousal, Valence in the continuous range [1.0, 9.0]	Classification: Low/High Arousal Low/High Valence
SEED [36]	EEG	-Determined from the study's authors -Fixed for all participants	3 discrete classes: Negative, Neutral, Positive	Classification: Negative, Neutral, Positive
AffectNet [20]	Images	-Determined by multiple external annotators	8 discrete classes* and Arousal, Valence in the continuous range [-1.0, 1.0]	Regression: Arousal, Valence

TABLE II: Details regarding the affective annotations and evaluation tasks on the datasets used in our work.
* We do not use the discrete class labels of AffectNet in any stage of our work.

The discrete class annotations of SEED are traditionally treated as being nominal, ignoring the evident ordinality. The classes of SEED practically correspond to three ordered levels of valence, therefore inferring ordinal relations between them is plausible. We adopt the convention that the "Positive" class corresponds to higher valence compared to "Neutral" and "Negative", and that the "Neutral" class corresponds to higher valence compared to "Negative". These ordinal relations that are used on SEED dataset are shown in Table III.

y_1 \ y_2	**Negative**	**Neutral**	**Positive**
Negative	$x_1 \sim x_2$	$x_1 \prec x_2$	$x_1 \prec x_2$
Neutral	$x_1 \succ x_2$	$x_1 \sim x_2$	$x_1 \prec x_2$
Positive	$x_1 \succ x_2$	$x_1 \succ x_2$	$x_1 \sim x_2$

TABLE III: The ordinal relations that are adopted in our work, for the categorical labels of SEED dataset to be rendered useful in the pairwise ranking task.

AffectNet dataset: AffectNet [20] is a dataset of facial images annotated both in terms of discrete facial expression classes and continuous Arousal/Valence in the range of $[-1, 1]$. There are 280K images in the training set, and 4K images in the validation set. We do not make use of the categorical labels in any way, and our end-goal task is the regression of arousal and valence, hence the regression head has two outputs.

Considering the dataset of AffectNet in the context of regression, the involvement of multiple annotators with subjective perception biases, presents a challenging case for ranking. Forming sample pairs to perform ordinal comparisons with respect to their ratings, is typically done on samples from a single human annotator. However, AffectNet does not provide information to establish correspondences between ratings and individual annotators. Thus in our approach we rank pairs of samples from unknown annotators and deal with additional sources of label noise. Ranking is applied following the rules of Table I.

B. Experiments on DEAP and SEED

EEG data preparation: To perform training on DEAP and SEED, we represent each input sample in the form of a feature vector. Among the most well-established EEG signal features for emotion recognition, are Power Spectral Density (PSD), Power Spectral Asymmetry (PSA) and Differential Entropy (DE). For each electrode's signal, these features are computed in a specific frequency band and for a short time window (2 seconds on DEAP, 1 second on SEED). There are 5 frequency bands that are commonly used for feature extraction, namely theta band ($4-8$ Hz), alpha band ($8-12$ Hz), slow alpha band ($8-10$ Hz), beta band ($12-30$ Hz) and gamma band ($30-45$ Hz). PSD features characterize the spectral content of each signal, while PSA features measure the asymmetric hemisphere activation occuring in the brain through pairs of laterally corresponding/symmetric electrodes. We compute PSD and PSA as in [10], using the method of Welch [31]. The DE features measure the complexity of the signal across time [5]. On DEAP dataset, we use the PSD and PSA features, concatenating their feature vectors. On SEED, we use the precomputed DE features, provided by [36]. On both datasets, to discard features of negligible discriminability, feature selection is applied using Fisher's linear discriminant, similarly to [10], keeping the 100 most discriminative features. Afterwards, a zero-mean and unit-variance normalization procedure is applied on each of the remaining features, using the statistics of the train set.

Training details for DEAP, SEED: Training is done for 20 epochs with a batch size of 40, using a Stochastic Gradient Descent (SGD) optimizer, learning rate $lr = 0.001$, momentum $m = 0.9$ and weight decay equal to 5e-4. For DEAP dataset, the ordinal ranking operation is performed setting $\epsilon = 0.25$ and following Table I. The training process is a subject-dependent 10-fold cross validation. For each subject the 40 available trials are split into 10 folds (each fold containing 4 trials), keeping 9 folds as the train set and 1 fold as the test set. For SEED dataset, the ordinal ranking operation is performed following Table III. The training process is subject-dependent and the train-test splits are done in the same way with [36]. On both datasets, evaluation is done by computing the classification accuracy and F1 score. Especially on DEAP where there is significant class imbalance, the F1 score is a more representative measure of model performance.

Our experiments explore the impact of joint training on the model classification performance. As a baseline method, a plain MLP network (with 2 FC layers in its embedding module and 1 FC classification layer) is trained only on the classification task. In our case, we train PRNet jointly on the classification and ranking tasks. From the results of Table IV

and Table V, we can see that joint training improves the accuracy and F1 score both on the dataset of DEAP and SEED. Considering the F1 scores, the performance improvement of the proposed method over the baseline is statistically significant on DEAP ($p < 0.01$ for both arousal and valence), but not on SEED ($p = 0.058$).

Model	Arousal		Valence	
	Acc.	F1	Acc.	F1
Classification loss	60.49	51.94	57.69	54.61
Proposed method: Classification + ranking loss	**60.60**	**53.25***	**58.42**	**55.57***

TABLE IV: Accuracy (%) and F1 score on DEAP dataset. Stars indicate statistical significance of the F1-score distribution over subjects, according to Student's t-test ($* = p < 0.05$)

Model	3-class problem	
	Acc.	F1
Classification loss	74.80	72.79
Proposed method: Classification + ranking loss	**76.98**	**75.51**

TABLE V: Accuracy (%) and F1 score on SEED dataset.

The results verify our motivation of forming and learning pairwise relations utilising the available affective annotations. On DEAP, we notice that collapsing fine-grained affective rating information into discrete classes, is harmful for the training process. Incorporating the ranking supervision through the ordinal relation labels derived by the original continuous ratings, we boost the performance of our model. Similarly, the fact that our approach considers the ordinality of the classes on SEED, shows that our method can be beneficial even in cases where the original annotations are discrete.

C. Experiments on AffectNet

Visual data preparation: Training on AffectNet dataset is performed using images as inputs. Data augmentation is performed during training, through rotation ($\pm 30°$), zoom ($\pm 15\%$), horizontal flipping and brightness/contrast changes. The images are finally resized in a 96×96 size and their pixel values are normalized in the $[0, 1]$ range by dividing with 255.

Training details for AffectNet: Training is done for 50K steps with a batch size of 128. A Stochastic Gradient Descent optimizer is used, with learning rate $lr = 0.001$, momentum $m = 0.9$ and weight decay equal to 1e-4. The ordinal ranking operation is performed setting $\epsilon = 0.15$ and following Table I. A plain CNN network is trained only on the regression task, serving as our baseline model. Joint training on regression and ranking is performed with PRNet and the results are shown on Table VI. We use the evaluation measures of Root Mean Square Error (RMSE), Pearson Correlation Coefficient (PCC) and Concordance Correlation Coefficient (CCC), as defined in [20]. By design, the RMSE measure is sensitive to outliers. Also, RMSE does not reflect the covariance of data while the PCC measure does so. Compared to PCC, CCC differs in the sense that it scales the correlation of the two distributions with the distance between their means.

Model	Arousal			Valence		
	RMSE	CCC	PCC	RMSE	CCC	PCC
MSE loss	0.365	0.405	0.488	0.430	0.527	0.577
Proposed: MSE+rank	**0.350**	**0.461**	**0.529**	**0.409**	**0.567**	**0.605**

TABLE VI: Root Mean Square Error (RMSE ↓), Concordance Correlation Coefficient (CCC ↑) and Pearson Correlation Coefficient (PCC ↑) on the validation set of AffectNet.
↓: lower means better, ↑: higher means better.

The results show that adding the pairwise ranking loss as an extra supervision during training, has beneficial effects for both arousal and valence on all measures. One interesting observation of the results, is that incorporating the ranking task does not adversely affect the model's performance on the RMSE measure. The original MSE loss that is used to train the network on the regression task, explicitly optimizes the performance on the measure of RMSE. The improvement obtained on RMSE by jointly training on the regression and ranking task, indicates that the potential merits of ranking approaches are not restricted on particular measures.

To compare our method against RankNet [2] on the AffectNet dataset, we select the same architecture as our baseline CNN and train it using the method of [2]. We evaluate the model using PCC measure between the sample-wise preference scores and the ground truth arousal/valence ratings. The results (PCC=0.517 for arousal, PCC=0.580 for valence) are inferior to those obtained with our proposed method. This shows the importance of performing comparisons using trainable functions instead of fixed operations.

D. Ablation studies

To get deeper insights about the incorporation of the ranking task in the training process, we perform ablation studies using the dataset of AffectNet.

Impact of ranking loss: In this experiment, we investigate the impact of the coefficient α that is used to weight the contribution of the ranking loss to the total objective. In Table VII we report the performance of PRNet on the validation set of AffectNet, for different values of α. It can be seen that increasing the contribution of the ranking loss in the total optimization objective, can yield boosts especially in terms of valence estimation. In the case of arousal the benefits are smaller, indicating that the value of the ranking coefficient α should be estimated on a case-dependent basis.

Impact of rank tolerance: As explained in Section III, the hyperparameter of rank tolerance ϵ defines the maximum allowed distance between two labels on the rating scale, for their samples to be considered as similar. Changing the value of ϵ leads to the reformulation of the ordinal ranking labels that correspond to pairs of samples. The rank tolerance directly affects the sensitivity of the ranking module on discriminating pairs of higher/lower ratings. The results are shown in Table VIII.

Ranking coefficient	Arousal			Valence		
	RMSE	CCC	PCC	RMSE	CCC	PCC
Baseline ($\alpha = 0$)	0.365	0.405	0.488	0.430	0.527	0.577
$\alpha = 0.5$	0.351	0.483	**0.553**	0.407	0.568	0.604
$\alpha = 1.0$	0.350	0.461	0.529	0.409	0.567	0.605
$\alpha = 2.0$	0.351	0.484	0.540	0.408	0.565	0.606
$\alpha = 3.0$	0.350	0.461	0.529	0.409	0.567	0.605
$\alpha = 4.0$	0.352	0.487	0.548	**0.385**	**0.609**	**0.638**
$\alpha = 5.0$	**0.343**	**0.494**	0.550	0.397	0.602	0.622
$\alpha = 6.0$	0.347	0.490	0.547	0.404	0.581	0.612

TABLE VII: Ablation study on the impact of the ranking coefficient α on the model performance.

Rank tolerance	Arousal			Valence		
	RMSE	CCC	PCC	RMSE	CCC	PCC
Baseline (no rank)	0.365	0.405	0.488	0.430	0.527	0.577
$\epsilon = 0.05$	0.347	0.472	0.513	0.400	0.579	0.603
$\epsilon = 0.10$	0.357	0.461	0.536	0.399	0.585	0.620
$\epsilon = 0.15$	0.350	0.461	0.529	0.409	0.567	0.605
$\epsilon = 0.20$	**0.347**	0.459	0.509	0.405	0.562	0.612
$\epsilon = 0.25$	0.353	**0.478**	**0.542**	0.406	0.579	0.601
$\epsilon = 0.30$	0.361	0.461	0.537	**0.393**	**0.596**	**0.629**
$\epsilon = 0.35$	0.352	0.465	0.528	0.424	0.544	0.597

TABLE VIII: Ablation study on the impact of the rank tolerance hyperparameter ϵ.

The selection of hyperparameter ϵ has a noticeable impact on the performance of models trained with ranking loss. Specifically, the largest gains on the estimation of valence are achieved by choosing relatively high values of rank tolerance ($\epsilon = 0.30$). Regarding arousal, the optimum performance in terms of RMSE is achieved when $\epsilon = 0.20$, while for CCC/PCC it is achieved when $\epsilon = 0.25$. We can see that there are consistent improvements over the baseline, and therefore the framework is robust to the precise value of ϵ and gives benefits even in the case of large values of ϵ. This is consistent with the benefits in SEED dataset where the quantization of the label space is rather rough (i.e., not fine-grained).

Generalization benefits of ranking: In this experiment, we investigate the performance of models that are trained on a training set that does not contain labels at the ends of the distribution (i.e., very high or very low) and evaluated on the original validation set. More specifically, the original label space of the continuous annotations for arousal and valence, covers the range $[-1, 1]$. We introduce a cut-off hyperparameter c ($0 < c \leq 1$) and discard samples from the training set, by restricting the label space of the training samples so as to cover the range $[-c, c]$. That is, we discard a training sample x having labels (y_a, y_v), if either of the conditions $|y_\text{a}| > c$ and $|y_\text{v}| > c$ holds true. Lower values of c lead to smaller training set and narrower training label space, rendering the generalization of the trained model on validation samples that cover the original label space (i.e., $[-1, 1]$) a challenging task. We compare baseline models that are trained on the regression task, with models that are trained jointly on regression and ranking. It is expected that lower values of c deteriorate more the validation performance of *all* trained models, as they are evaluated on samples belonging to an increasingly large unseen label space. The results are shown in Table IX.

Model, c	Arousal			Valence		
	RMSE	CCC	PCC	RMSE	CCC	PCC
Baseline c=0.4	0.521	0.004	0.044	0.468	0.160	0.402
Proposed c=0.4	**0.497**	**0.035**	**0.154**	**0.440**	**0.247**	**0.457**
Baseline c=0.5	0.491	0.023	0.125	0.462	0.300	0.456
Proposed c=0.5	**0.471**	**0.070**	**0.228**	**0.441**	**0.355**	**0.498**
Baseline c=0.6	0.452	0.079	0.276	0.460	0.357	0.482
Proposed c=0.6	**0.424**	**0.176**	**0.359**	**0.434**	**0.421**	**0.536**
Baseline full	0.365	0.405	0.488	0.430	0.527	0.577
Proposed full	**0.350**	**0.461**	**0.529**	**0.409**	**0.567**	**0.605**

TABLE IX: Ablation study on the generalization capabilities of models trained with a restricted label space. Evaluation is done on the full validation set of AffectNet (i.e. validation labels of the entire range $[-1, 1]$).

We observe that shrinking the training label space is less harmful to the performance of the models that are trained jointly on regression and ranking, as models that are trained on plain regression fail to generalize. This is supported by the increasingly large performance gaps, especially for arousal but also for valence, as c takes lower values.

Discussion: We applied our method on datasets where the original affective labels were converted into pairwise ranking labels. Transforming affective ratings into ordinal annotations, is a more controllable process when done on a per-subject basis [16] (i.e., comparing labels of a single annotator). On our pairwise ranking labels formed for the AffectNet dataset, we went beyond this practice, by comparing ratings of different annotators. Furthermore, the self-assessment ratings of each participant on DEAP dataset, were done in an one-hour long session. Hence, establishing pairwise comparisons between temporally distant ratings may result in less reliable ranking labels, due to the potentially different context within which the ratings were reported.

V. Conclusion

The findings of our work highlight that exploring the ordinality of emotions through deep neural networks that accomodate pairwise ranking comparisons, is beneficial for affect recognition models. The proposed method is evaluated on neurophysiological and visual data with diverse affective annotation processes, showing consistent performance gains. The performed ablation studies shed light on various aspects of the ordinal ranking task. We believe that our study provides a promising direction on training robust emotion recognition models, through tasks that abide to the ordinal nature of emotions.

REFERENCES

[1] T. Ayral, M. Pedersoli, S. Bacon, and E. Granger. Temporal stochastic softmax for 3d cnns: An application in facial expression recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3029–3038, 2021.

[2] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[3] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[4] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, Jan 2010.

[5] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 81–84. IEEE, 2013.

[6] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

[7] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.

[8] J. Fürnkranz and E. Hüllermeier. *Preference learning*. Springer, 2010.

[9] H. Helson. Adaptation-level theory: an experimental and systematic approach to behavior. 1964.

[10] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[11] D. Kollias, S. Cheng, E. Ververas, I. Kotsia, and S. Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, pages 1–30, 2020.

[12] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.

[13] W. Lin, C. Li, and S. Sun. Deep convolutional neural network for emotion recognition using eeg and peripheral physiological signal. In *International Conference on Image and Graphics*, pages 385–394. Springer, 2017.

[14] P. Lopes, G. N. Yannakakis, and A. Liapis. Ranktrace: Relative and unbounded affect annotation. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 158–163. IEEE, 2017.

[15] F. Ma, B. Sun, and S. Li. Robust facial expression recognition with convolutional visual transformers. *arXiv preprint arXiv:2103.16854*, 2021.

[16] H. P. Martinez, G. N. Yannakakis, and J. Hallam. Don't classify ratings of affect; rank them! *IEEE transactions on affective computing*, 5(3):314–326, 2014.

[17] D. Melhart, A. Liapis, and G. N. Yannakakis. Pagan: Video affect annotation made easy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 130–136. IEEE, 2019.

[18] D. Melhart, K. Sfikas, G. Giannakakis, and G. Y. A. Liapis. A study on affect model validity: Nominal vs ordinal labels. In *Workshop on Artificial Intelligence in Affective Computing*, pages 27–34. PMLR, 2020.

[19] A. Metallinou and S. Narayanan. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

[20] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, pages 1–1, 2018.

[21] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari. Survey on emotional body gesture recognition. *IEEE transactions on affective computing*, 2018.

[22] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 579–595, 2018.

[23] S. Parthasarathy, R. Cowie, and C. Busso. Using agreement on direction of change to build rank-based emotion classifiers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2108–2121, 2016.

[24] S. Parthasarathy, R. Lotfian, and C. Busso. Ranking emotional attributes with deep neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4995–4999. IEEE, 2017.

[25] R. Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

[26] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[27] J. A. Russell and L. F. Barrett. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805, 1999.

[28] J. A. Russell and U. F. Lanius. Adaptation level and the affective appraisal of environments. *Journal of Environmental Psychology*, 4(2):119–135, 1984.

[29] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2011.

[30] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[31] P. Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

[32] G. N. Yannakakis, R. Cowie, and C. Busso. The ordinal nature of emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 248–255. IEEE, 2017.

[33] G. N. Yannakakis and H. P. Martinez. Grounding truth via ordinal annotation. In *2015 international conference on affective computing and intelligent interaction (ACII)*, pages 574–580. IEEE, 2015.

[34] G. N. Yannakakis and H. P. Martínez. Ratings are overrated! *Frontiers in ICT*, 2:13, 2015.

[35] S. Zhang, Z. Huang, D. P. Paudel, and L. Van Gool. Facial emotion recognition with noisy multi-task annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 21–31, 2021.

[36] W.-L. Zheng and B.-L. Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.