Deep Explanatory Polytomous Item-Response Model for Predicting Idiosyncratic Affective Ratings

Yan Zhou, Tsukasa Ishigaki
Graduate School of Economics and Management
Tohoku University
Miyagi, Japan
yan.zhou01@icloud.com, isgk@tohoku.ac.jp

Shiro Kumano
NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
Kanagawa, Japan
kumano@ieee.org

***Abstract*—Towards explainable affective computing (XAC), researchers have invested considerable effort into post hoc approaches and reverse engineering to seek explanations for deep learning models. However, alternative, intrinsic approaches that aim to build inherently interpretable models by restricting their complexity are yet to be widely explored. In this study, we integrate an explanatory polytomous item response model that provides a well-established psychological interpretation for ordinal scales with deep neural networks to realize high prediction performance and good result interpretability. We conducted an experiment on a growing task (i.e., predicting the idiosyncratic perception of emotional faces of an individual); as expected theoretically, the topmost parameters of our model demonstrated strong correlations with those of the corresponding ordinal item response model: r = 0.928 to 1.00. Our proposed intrinsic approach can used as a complementary framework for post-hoc methods in XAC to coach and support human social interactions.**

***Index Terms*—ordinal model, item-response theory, perceived emotion, affect dimension, explainable AI**

I. Introduction

Among the goals of affective computing (AC): building machines that recognize human emotions, behave emotionally, and have emotions [31], emotion recognition continues to be a popular interest in the community. Given the communicative view of human emotional expression [35], the research community targets two types of subjective emotions to recognize: emotions felt by the self (felt emotions) and emotions perceived by others (perceived emotions) [9], [38]. Both can be represented as basic emotions (e.g., happiness, anger, surprise) or emotional dimensions (e.g., valence and arousal); however, there is growing evidence which suggests ordinal representations may more accurately reflect the underlying experience [51]. A major issue for perceived emotions is determining the ground truth, and several previous studies assumed a single ground truth by taking the majority or mean of the observers' judgments to eliminate individual differences.

Considering the idiosyncratic perceptions of the observer, recent trends in the AC community include treating uncertainty as soft classification to maintain labels [42] or to build personalized recognizers that predict how an individual perceives a target stimulus [17], [58], although limited, unlike personalized felt emotion classification [8], [54]. To this end, deep neural networks play a tremendous role in enhancing performance [22].

Another trend, in the larger domain of artificial intelligence (AI), is that many researchers are seeking to develop *explainable AI* (XAI) [1], [3] for applications that require accountability for the machine's decisions to combat the performance-interpretability tradeoff [33]. Thus far, there are two XAI approaches: post hoc and intrinsic [1]. On the one hand, the post-hoc approach is based on a reverse engineering process that provides the required explanations without altering or knowing the inner workings of the original model. On the other hand, the intrinsic method constructs the model to be inherently interpretable or self-explanatory, which restricts its complexity. Post-hoc approaches are dominant in deep neural networks [33]. For example, saliency maps are considered explanatory for determining the parts of an image that are focused on by the classifier. However, this does not cover information regarding how the model uses the relevant information. Thus, there is a need to enhance interpretability when using intrinsic approaches, even at the risk of performance [33].

As one of the most central components of positive computing and subjective well-being [5], AC systems need to explain how they reach their prediction/decision and how they may affect the well-being of the user. For instance, systems that help and guide people to enhance their emotional interactions with others, such as building rapport, can be easily anticipated. Broader examples include personnel selection, education, advertising, autonomous driving (e.g., trolley problem), and court trials, as summarized in [15]. For such applications, not only rationality but also affect play key roles, and machines should account for their decisions. Therefore, we refer to AC equipped

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

with explainability as *explainable affective computing (XAC)*, which is a term that has hardly been used, besides in [40].

The realization of psychological interpretability with high prediction performance remains a noted challenge in idiosyncratic perceived emotion recognition. For example, in [17], a naïve Bayes model that separates the effect of observers and the stimulus based on their conditional independence assumption was proposed. However, their model is not psychologically parameterized, and therefore, it is challenging to provide psychological explanations. Further, their model required handcrafted emotion features, but it did not demonstrate significant prediction accuracy. Another study [58] achieved high performance in predicting an individual's perception of emotion images using rolling multitask hypergraph learning, which jointly combines multiple factors such as visual content and the metadata of images. Their model is not easy to interpret because of the complexity of the model structure.

A psychological theory such as the explanatory item response (IR) theory [46] has the potential to overcome the above-mentioned challenge. This theory comprises two layers: The higher layer is the base descriptive IR model, which assumes a stochastic process to describe how an individual responds to an observed item. The discrete IR theory assumes that latent parameters explain the logit of the probability that a person will provide a score to an item. The lower layer regresses the latent variables from auxiliary information about respondents or items, which enables the model to predict responses to unknown items or respondents. Researchers prepare easy-to-interpret auxiliary information to allow the explanatory IR models to be intrinsically interpretable in both the base IR and explanatory latent regression layers. However, in the AI community, it is well known that such handcrafted features degrade the predictive performance of the model [22]. In addition, it is plausible to employ it as a latent regressor only in the lower layer and maintain the interpretability of the IR layer, while exploiting the high feature extraction performance of the deep neural network.

In this paper, we propose a model for predicting the individual's perception of facial emotion images on an ordinal scale by incorporating a convolutional neural network (CNN) as the latent regression layer of the IR model to extract visual features. We apply the consistent rank logits (CORAL) framework [6] to ensure the ordinal consistency of the final results, which is arguably required for valid psychological interpretability. We theoretically and experimentally demonstrate that the top layer parameters of the proposed model are interpretable as psychologically established IR parameters while maintaining the performance of the CNN in the bottom layer to regress the latent IR parameters.

The contributions of this paper are two-fold:

1) This is the first AC study that integrates deep neural networks with explanatory polytomous IR theory to combat the performance-interpretability tradeoff.
2) We theoretically and experimentally demonstrate that CORAL can be combined with several types of IR models, which helps maintain the rank consistency of the response ratings.

II. Related Work

Recently, ordinal annotation has received substantial attention in the AC community [51] as an alternative approach to overcome the validity and reliability issues of nominal (e.g., basic emotion categories) and interval (e.g., valence-arousal dimension) approaches [26], [37]. The primary criticism received by the nominal approach is that emotions cannot be assigned to a single category in reality [39]. The interval approach aims to solve the issue using emotional dimensions; however, individual differences of interval labels are large [11] partly because of the limited ability of humans to express their preferences directly in terms of values. Recent AC studies have demonstrated improved performance over interval or nominal approaches as another alternative to overcome these issues [32]. However, modeling the subjective ordinal ratings of an individual in an explanatory way is not extensively explored.

An ordinal deep neural network is a popular topic in the AI community, e.g., in age estimation [6], [28]. In [28], an ordinal regression problem was converted into a series of binary classification subtasks. However, the validity and interpretability of the model are questionable because it cannot guarantee the rank consistency of the ordinal ratings. This issue was addressed in [6]. Their rank-consistent ordinal CNN (CORAL), which theoretically guarantees rank consistency, inspired us to combine their model with specific types of polytomous IR models for combating the performance-interpretability tradeoff in the problem of idiosyncratic perceived emotion recognition.

Several studies have proposed combining deep learning with *dichotomous* or binary IR theory [7], [47], [53]. In [7], multiple deep learning models were used to regress the parameters of the IR model to predict if a student answers the question correctly. In [53], a dichotomous IR model was integrated with deep learning to estimate the temporal change in the knowledge states of a student. In [47], an IR model was placed in the middle of a deep neural network, instead of on the top, where it faces difficulties with interpreting the output of the model using the IR model. Unlike these previous studies, our study focuses on different technical aspects and application domains. First, these studies used dichotomous responses, whereas we handled *polytomous (more specifically, ordinal)* affective ratings by incorporating the CORAL framework with the IR theory. Second, the previous deep IR models were trained in two steps from deep neural networks to IR models, whereas our model is trainable end-to-end.

III. Proposed Model

A. Consistency-reserved Deep Neural Network for Ordinal Regression

CORAL [6] is a framework for incorporating neural networks with ordinal regression to predict a target label (in our case, the emotional rating for a facial image) y^{org} in an ordered set $Y = \{r_1, \cdots, r_S | r_1 \prec \cdots \prec r_S, S \in \mathbb{N}\}$. The label extension transforms the rating y^{org} into $S-1$ binary labels

$[y_i^{(1)}, ..., y_i^{(S-1)}]^\intercal$, each of which indicates whether y_i exceeds rank s, i.e., $y_i^{(s)} = \mathbb{1}\{y_i \succ r_s\}$. The indicator function $\mathbb{1}\{\cdot\}$ is one if the inner condition is true, and zero otherwise. For example, let $y^{org} = r_3$ and $S = 5$; then, we have $y_i^{(1)} = 1$, $y_i^{(2)} = 1$, $y_i^{(3)} = 0$, and $y_i^{(4)} = 0$.

Let the output of the penultimate layer be denoted as $H(\mathbf{x}_i, \mathbf{W})$. Here, $\mathbf{W}$ denotes the parameters of the neural network excluding the bias units of the final layer. The CORAL framework ensures prediction consistency by allowing $H(\mathbf{x}_i, \mathbf{W})$ to share the same weight with all nodes in the final output layer; this helps minimize model complexity by downsizing the number of parameters. However, each node has an independent bias parameter κ_s. Let the output of the s-th binary classifier for an input image $\mathbf{x}_i$ be denoted by $g_s(\mathbf{x}_i) \in \{0, 1\}$. The value of $g_s(\mathbf{x}_i)$ is calculated from $g_s(\mathbf{x}_i) = \mathbb{1}\{P(\hat{y}_i^{(s)}) > 0.5\}$. Here, $P(\hat{y}_i^{(s)})$ indicates the probability that $\hat{y}_i$ exceeds r_s, i.e., $P(\hat{y}_i \succ r_s)$.

The predicted rank $\hat{y}_i$ for $\mathbf{x}_i$ is then obtained as $\hat{y}_i = r_q$, where $q = 1 + \sum_{s=1}^{S-1} g_s(\mathbf{x}_i)$. The probability that the rank of the predicted item is likely to exceed the rank r_s is defined as

$$P(\hat{y}_i^{(s)}) = P(\hat{y}_i \succ r_s) = \sigma(H(\mathbf{x}_i, \mathbf{W}) + \kappa_s), \tag{1}$$

where κ_s denotes a score-specific bias unit and $\sigma(z) = 1/(1 + exp(-z))$ represents a logistic sigmoid function. This is a cumulative logistic form where the cumulative probability on the left-hand side is linked with the linear component, i.e., the argument of the logistic sigmoid function on the right-hand side.

CORAL minimizes the following cross-entropy loss function of $S-1$ binary classifiers:

$$\begin{aligned} L(\mathbf{W}, \kappa) = -\sum_{i=1}^{N_i} \sum_{s=1}^{S-1} [&\log(\sigma(H(\mathbf{x}_i, \mathbf{W}) + \kappa_s)) y_i^{(s)} \\ &+ \log(1 - \sigma(H(\mathbf{x}_i, \mathbf{W}) + \kappa_s))(1 - y_i^{(s)})]. \end{aligned} \tag{2}$$

Weight sharing in the final output layer guarantees that the global optimum satisfies the rank consistency, i.e., $S-1$ cumulative probabilities are monotonically decreasing: $P(y_i \succ r_1) \geq P(y_i \succ r_2) \geq \ldots \geq P(y_i \succ r_{S-1})$. In addition, the probability of each rank can be obtained by taking the difference between adjacent cumulative probabilities $P(y_i = r_s) = P(y_i \succ r_{s-1}) - P(y_i \succ r_s)$, except for $P(y_i = r_1) = 1 - P(y_i \succ r_1)$ and $P(y_i = r_S) = P(y_i \succ r_{S-1})$. The estimated rank is determined to have the maximum probability.

B. Base Explanatory IR Model

The main idea of this study is combining CORAL with the explanatory IR polytomous theory to handle the heterogeneity of the respondents. Among the several types of IR models, we select the one-parameter rating-scale graded response model (1P-RS-GRM) [4]. 1P-RS-GRM is expressed as $P(\hat{y}_{ij} \succ r_s) = \sigma(\theta_j - \beta_i - \kappa_s)$. Like in other IR models, θ_j, β_i, and κ_s denote the j-th respondent (ability) parameter, i-th item (difficulty) parameter, and s-th threshold location parameter of the rating scale (called the rating scale parameter), respectively. θ denotes the sensitivity of the respondent to the emotional dimension. A larger θ indicates that the person is more sensitive to the dimension and thus tends to assign greater ratings [45]. β denotes the difficulty of the item to obtain a greater rating. The threshold parameters κ represent the $(K-1)$ threshold locations, each of which determines whether the perceptual representation $\theta - \beta$ exceeds threshold κ. Further, it plays a similar role as the item difficulty parameter; however, it is score-specific and item-independent.

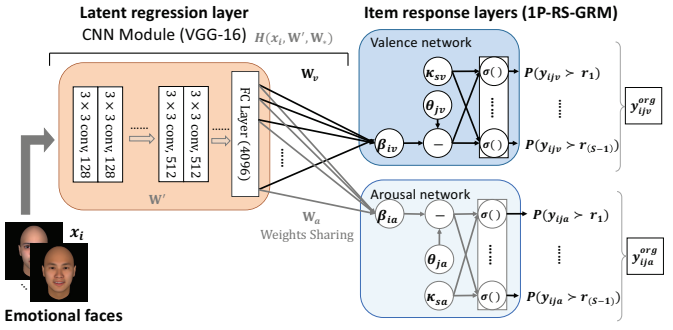

Fig. 1. Model architecture of the proposed CORAL-RS-GRM. It comprises two parts: i) On the left (bottom layer) side, one shared CNN regresses the item parameter β from the input item image for both valence and arousal dimensions. The input of β is the final layer of the CNN module. ii) On the right (top layer) side, the item-response model (more specifically, 1P-RS-GRM) layers generate affective rating y by considering θ and the rating scale parameter κ for both dimensions separately. Unlike β, θ and κ have no input. The output is the rating in the extended form. In standard neural network terminology, β is a hidden layer (with a single node), whereas θ and κ are bias parameters (but respondent- and score-specific, respectively). Therefore, the entire model, including the IR layer, can be learned using ordinary gradient-based algorithms.

In explanatory IR models [49], β_i is regressed using auxiliary information about the item property, $\mathbf{x}_i$, i.e., $\beta_i = f(\mathbf{x}_i)$. Function f is frequently a linear regression; however, it can be in any form. The explanatory 1P-RS-GRM is expressed as

$$P(\hat{y}_{ij} \succ r_s) = \sigma(\theta_j - f(\mathbf{x}_i) - \kappa_s). \tag{3}$$

Here, notice the similarity between (1) and (3): First, the right-hand side of both equations includes a score-specific bias in CORAL and the threshold parameter in 1P-RS-GRM, respectively, as represented by κ_s. Second, $H(\mathbf{x}_i, \mathbf{W})$ in (1) corresponds to $f(\mathbf{x}_i)$ in (3), while (3) includes θ_j. Thus, we built a deep explanatory 1P-RS-GRM by adding θ_j to the CORAL framework.

C. Proposed CORAL-RS-GRM

We integrate CORAL with 1P-RS-GRM (CORAL-RS-GRM) as illustrated in Fig. 1. Our model defines the cumulative probability of the rating of respondent j to item i exceeding rank r_s as

$$P(\hat{y}_{ij} \succ r_s) = \sigma(\theta_j - H(\mathbf{x}_i, \mathbf{W}) - \kappa_s) \tag{4}$$

The main advantage of the proposed model is that we can interpret θ_j, $H(\mathbf{x}_i, \mathbf{W})$, and κ_s as the respondent parameter, item difficulty parameter, and rating scale parameter, respectively, like that in the IR theory.

Following the recent neural-network-based emotion estimation studies [30], [56], we apply the multitask learning fashion. This is more advantageous than single-task learning, particularly when training data are limited by sharing low-level representations for multiple tasks. As shown in Fig. 1, the model has two-branched networks from the CNN module to the valence and arousal dimensions separately. They share the same CNN module as the latent regression layer; however, they have their own IR layers: $\mathbf{W} = \{\mathbf{W}', \mathbf{W}_*\}$. $\mathbf{W}_*$ denote the weights shared with all nodes in the final output layer of each branched network as imposed by the CORAL framework. $\mathbf{W}'$ denotes the rest of the CNN module parameters shared by the two sub-networks. The two sub-networks have their own IR model parameters, θ_{j*}, β_{i*} (dependent on item i, $\mathbf{W}'$, and $\mathbf{W}_*$), and κ_{s*}, where * denotes an emotional dimension (e.g., $* \in \{\text{valence}, \text{arousal}\}$.

We minimize the total loss function

$$L_{\text{total}}(\mathbf{W}, \theta, \kappa) = \sum_{*} L(\mathbf{W}', \mathbf{W}_*, \theta_*, \kappa_*), \tag{5}$$

$$\begin{aligned} &L(\mathbf{W}', \mathbf{W}_*, \theta_*, \kappa_*) \\ &= \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \sum_{s=1}^{S-1} [\log(\sigma(\theta_{j*} - H(\mathbf{x}_i, \mathbf{W}', \mathbf{W}_*) - \kappa_{s*})) y_{ij*}^{(k)} \\ &\quad + \log(1 - \sigma(\theta_{j*} - H(\mathbf{x}_i, \mathbf{W}', \mathbf{W}_*) - \kappa_{s*}))(1 - y_{ij*}^{(k)})], \end{aligned}$$

where N_i and N_j denote the number of items and respondents, respectively.

IV. Evaluation Experiments

A. Experimental Data

We used the dataset collected in [18] to evaluate our model. The dataset includes ratings given by various persons to emotional faces. Although there are many open image datasets such as CK+ [24], MMI [29], AffectNet [27], and IAPS [21], there are very limited datasets that are publicly available and include the ratings of individual respondents, as summarized in [57].

A total of 50 respondents, who were all Japanese students in domestic universities, were asked to judge 120 emotional face images. The rating process followed a blocked design: one block was used for valence judgment, and the other was used for arousal. The rating was a forced-choice on a five-point scale in each block (i.e., $S = 5$): negative (1) versus positive (5) for valence, or low (1) versus high (5) for arousal. Each face was presented for 1,000 ms, following a fixation cross to the center of the screen for a duration of 500 ms. The stimulus exposure durations were short but sufficient, and they were used in studies such as [23]. Then, a valence or arousal scale was displayed until the participant selected one of the answers. The 120 faces include both single-category expressions out of eight categories (angry, disgust, fear, neutral, sad, surprised, and smiles with opened/closed mouth) and mixed expressions of pairs of the eight categories. All facial images were in the front view and under the same illumination. For details on the generation process of the 120 faces, see [18].

There were 150 trials per respondent in each block. In 30 of trials, an image was randomly selected from the 120 faces for the second rating. The pair of ratings were used only to calculate the test-retest reliability as an indicator of the upper bound of the prediction performance of the model. Therefore, we used 12,000 samples (120 items × 50 people × 2 dimensions) for the main analysis, and the remaining 3,000 samples were used only to calculate the test–retest reliability. Further, frequencies of the ratings were $\{1(\text{negative}) = 9\%, 2 = 32\%, 3 = 35\%, 4 = 20\%, 5(\text{negative}) = 4\%\}$ for valence, and $\{1(\text{low}) = 9\%, 2 = 25\%, 3 = 31\%, 4 = 28\%, 5(\text{high}) = 7\%\}$ for arousal.

B. Experimental Settings

We used the PyTorch framework to implement the proposed model. We used the Adam optimizer for training and set the learning rate = 2e-04, batch size = 10, and #epochs = 150. As the pre-trained model, we used the VGG-16-based facial recognition model trained for facial expression classification on the FER2013 dataset [14]; see [50]. Pre-training a model with multiple approximate datasets is referred to as the multi-stage pre-trained strategy, and it can discriminate ambiguous local features among similar categories. Further, this strategy mitigates the limited data issue [43]. We unfroze the last two convolutional layers of the pre-trained model in the fine-tuning on the main dataset because this yielded the best training loss when we unfroze the last one to four convolutional layers.

C. Baseline Models

We compared the proposed CORAL-RS-GRM with three IRT-based baseline models: Action-Unit-based latent regression 1P-RS-GRM (AU-RS-GRM, hereafter), and the base 1P-RS-GRM. The first two models are identical in their item-response layer but different in their latent regression layer, as summarized in Table I. Unlike CORAL-RS-GRM, AU-RS-GRM is interpretable in its latent regression layer. Therefore, this is a doubly interpretable model. The AUs are descriptors of the basic actions of an individual muscle or a group of muscles [13]. AUs are psychology-grounded, and thus, highly interpretable, such as the eyebrow lower (AU4) or lip corner puller (AU12). We detected AUs using OpenFace [2]. We detected the intensities of 17 AUs and the binary presence of 18 AUs. Then, we fed the 35 features into the latent regression layer of β, which is the first linear layer of the AU-RS-GRM. Therefore, the linear layer is essentially the linear regression. We created the rest of the layers in the same manner as the proposed model. The AU-based model can evaluate the performance-interpretability tradeoff compared with the proposed model.

The base IR model (1P-RS-GRM) has no latent-regression layer, which means that 1P-RS-GRM cannot handle unknown items and is therefore not applicable to the test sets. We consider the estimated parameters of 1P-RS-GRM to be the most accurate because this model is the least restricted among the three models in Table I. We implemented the models in the

TABLE I
DIFFERENCES BETWEEN THE PROPOSED AND IRT-BASED BASELINE MODELS

Model	IR layer	Latent-regression layer
1P-RS-GRM (Base IRM, least restricted)	Interpretable	-
CORAL-RS-GRM (Proposed)	Interpretable	High performance, Difficult to interpret
AU-RS-GRM (Doubly interpretable)	Interpretable	Interpretable, Low performance

Stan probabilistic programming language. We used the No-U-Turn Sampler (NUTS) to obtain accurate estimations of the IR model parameters. NUTS incurs a considerable computational cost; however, it seeks the global optimal compared to approximation-based methods such as gradient-based methods (e.g., Adam) or variational inference.

Being the least and strongest constrained structures among the three IR-based models listed in Table I, we expected that the base IR model and AU-RS-GRM would be the best and worst, respectively, in terms of training performance.

Further, we used RankSVM in pyplt [12] that takes AUs as input to predict ratings as a more general preference learning algorithm to better illustrate the performance of the proposed model. RankSVM was trained on each respondent and emotion dimension independently.

D. Evaluation Methods

We employed a leave-10%item-out cross-validation to evaluate the predictive performance and interpretability of the proposed model. We randomly split the dataset into ten equal parts item-wise, i.e., 90% of the 120 images (108 images) were used as the training set and the remaining 10% (12 images) were used as the test set. We used a set of several metrics in combination to capture multiple aspects of the results following [16].

We report the accuracy, mean absolute error (MAE), Somers' D, and normalized discounted cumulative gain (nDCG) to measure the training and test performance on the *ordinal, i.e. discrete* labels. Accuracy is commonly used in classification tasks. MAE is an error metric that has been widely used for interval/ratio scales. Somers' D is more suitable for label ranking tasks, and it ranges between -1 and 1. Somers' D compares the global ranks of all items, wherein -1 indicates that it is entirely inconsistent with the order of the label; 1 indicates that it is completely consistent. nDCG measures the gain score of an item based on its position in the result list, and the score ranges between 0 (worst) and 1 (best). We calculated the mean results of the four types of metrics on ten training and test sets. Further, we calculated the test–retest reliability in the same manner by approximating the upper bound of the prediction performance.

We calculated the Pearson's correlation coefficients (PCC) and MAE of the estimated IR parameters of the base (1P-RS-GRM) and proposed models to measure the extent of consistency and absolute errors, respectively for evaluating the similarity. Two metrics aim to validate the interpretability of the estimated IR parameters of the proposed model. The item parameters β were partially collected from each test set, and we obtained all estimations when the cross-validation was completed. However, because all respondent parameters θ and all rating scale parameters κ were obtained from each test set, we report their average values at the end.

TABLE II
TRAINING PERFORMANCE ON RATING ESTIMATION

(a) For training data

Model		Acc	MAE	S'D	nDCG
1P-RS-GRM (Base IRM, least restricted)	V	**0.650**	**0.383**	*0.637*	**1.000**
	A	*0.542*	**0.508**	*0.507*	**0.995**
CORAL-RS-GRM (Proposed)	V	0.575	*0.468*	**0.641**	*0.943*
	A	0.424	*0.651*	**0.519**	*0.953*
AU-RS-GRM (Doubly interpretable)	V	0.473	0.592	0.575	0.938
	A	0.375	0.741	0.433	0.939
RankSVM	V	*0.644*	1.777	-0.102	0.832
	A	**0.680**	1.972	-0.09	0.843

(b) For test data (i.e., unknown images)
Note: 1P-RS-GRM cannot be applied to test data because of the lack of latent regression layers.

		Acc	MAE	S'D	nDCG
CORAL-RS-GRM (Proposed)	V	**0.531**	**0.562**	**0.568**	**0.930**
	A	**0.402**	**0.736**	**0.418**	**0.920**
AU-RS-GRM (Doubly interpretable)	V	*0.441*	*0.590*	*0.526*	*0.927*
	A	*0.380*	*0.743*	*0.408*	*0.909*
RankSVM	V	0.218	1.776	-0.092	0.820
	A	0.246	1.971	-0.057	0.829
Test–Retest reliability	V	0.607	0.466	0.585	0.993
	A	0.495	0.635	0.495	0.990

Acc: accuracy, S'D: Somers'D, V: valence, A: arousal.

V. RESULTS

A. Performance

First, we compared the training performance of the four models to determine the basic characteristics; the comparison results are presented in Table II (a). The results are mostly consistent with our expectations. The base IR model (1P-RS-GRM) achieved the best accuracy of 0.650 for valence, MAE (0.383 for valence and 0.508 for arousal), and nDCG (1.000 and 0.995, respectively) among the three IR-based models. Our model (CORAL-RS-GRM) ranked second (accuracy = 0.575 and 0.424; MAE = 0.468 and 0.651; nDCG = 0.943 and 0.953), although it reached slightly higher Somers'D (0.641 for valence and 0.519 for arousal) than 1P-RS-GRM (0.637 and 0.507, respectively). Among the three IR-based models, the most explanatory model (AU-RS-GRM) was the worst, with the lowest accuracy (0.473 and 0.375) and the highest MAE (0.592 and 0.741). RankSVM achieved the highest accuracy of 0.680 for arousal and the second highest accuracy of 0.644 for valence. However, the other three metrics were found to be the worst among the four models.

Second, Table II (b) shows the test results of CORAL-RS-GRM, AU-RS-GRM, and RankSVM in the cross-validation. 1P-RS-GRM cannot be applied in the test sets because it does not have an explanatory layer. Our proposed model, CORAL-RS-GRM, outperformed AU-RS-GRM for two emotional dimensions of valence and arousal in all four metrics. The results validated our assumption that employing a CNN as an item latent regressor can help achieve higher generalizability. The accuracies of CORAL-RS-GRM were larger than those of AU-RS-GRM (20% for valence and 6% for arousal); however, the differences were smaller for MAE, Somers'D, and nDCG. Compared with the test–retest reliabilities as an approximated upper-bound performance, the proposed model showed good performance in terms of the four metrics.

On the test set, our model achieved the nDCG of 0.930 and 0.920 for the valence and arousal, respectively, whereas RankSVM yielded 0.820 and 0.829, and AU-RS-GRM achieved 0.927 and 0.909. In the training set, the results were slightly higher; however, the trend was the same. These results suggest the advantage of our model over RankSVM.However, we acknowledge that there is still room for improvement in RankSVM in terms of sharing item parameters across respondents and dimensions, as in our model.

Considering the imbalanced dataset, the acceptable baseline model that at least always returns the most frequent class achieves accuracies of 0.35 for valence and 0.31 for arousal. Moreover, providing the maximum frequencies and test–retest reliabilities (0.607 and 0.495, respectively) as rough lower and upper bounds, the accuracies of our model (0.531 and 0.402) were corrected to $70\% = (0.531-0.35)/(0.607-0.35)$ and $50\% = (0.402-0.31)/(0.495-0.31)$. These are promising results for such a challenging task.

Further, we confirmed that CORAL-RS-GRM satisfied the rank consistency for both two emotional dimensions. The estimated rating scale parameters increased monotonically: $\kappa = (-4.01, -0.86, 1.54, 4.22)$ for valence and $\kappa = (-2.93, -0.95, 0.84, 3.64)$ for arousal. The inconstant intervals for valence validate the need for the ordinal model, which is in line with the literature [51].

B. Interpretability

We examined the similarity of the estimated IR parameters between the proposed model and 1P-RS-GRM to evaluate whether the item-response parameters of our model can be interpreted in a similar way as those of the base IR model. The IR parameters of our model were obtained from the test results of the cross-validation for the proposed model (see IV-D), whereas those of 1P-RS-GRM were obtained from the training results (they were thus presumably closer to the ground truth), as explained in IV-C.

They were strongly correlated ($r > 0.92$), as summarized in Table III and shown in Fig. 2. The high similarities indicate that we can interpret the estimated parameters as in the base IR model. Moreover, although less insightful, Table III also includes MAE, which is an absolute error-based metric.

TABLE III
SIMILARITY OF ESTIMATED IR PARAMETERS BETWEEN PROPOSED MODEL AND BASE IR MODEL

		PCC (r)	MAE
Item parameter β	V	0.928	0.616
	A	0.935	0.429
Respondent parameter θ	V	0.967	0.819
	A	0.972	0.705
Rating scale parameter κ	V	1.000	0.245
	A	1.000	0.149

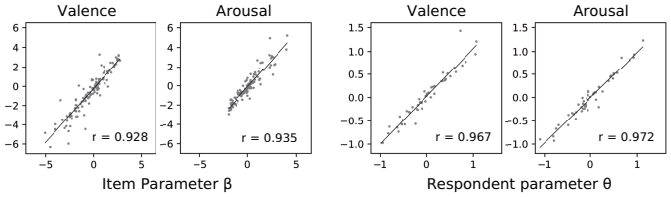

Fig. 2. Estimated parameters: Proposed method (horizontal) vs. Base IR model (vertical).

Fig. 3 shows a scatter plot of the estimated item parameters β in valence-arousal dimensions for qualitative evaluation. In Fig. 3, the expression categories used when generating faces on the modeler reasonably yielded category-specific clusters. For example, in terms of valence, angry faces had positive β values, which means that angry faces are likely to receive relatively inferior valence ratings (i.e., closer to rate 1 in our experiment's case). Smiling faces received negative values (i.e., likely to obtain relatively greater valence ratings, i.e., closer to rate 5 in our case). For arousal, angry faces received negative values (i.e., likely to receive relatively greater arousal ratings); sad and neutral faces received positive values (i.e., likely to receive relatively lower arousal scores). These results are in line with those in the psychological literature [25], [34].

Fig. 3 shows an asymmetric V-shaped relationship of β that has been reported in several emotion studies [20] (the slope of the regression line on the negative side is steeper than that on the positive side). The base IR model yielded a similar asymmetric V-shaped relationship. Both further validate the interpretability of the IR parameters in our model.

We also evaluated how the item parameters were regressed from AUs in the most interpretable model (AU-RS-GRM). We found large regression coefficients of AU intensity for valence: a positive coefficient (= 2.80) for AU12 and a negative coefficient (= -2.58) for AU4. These results are in line with those reported in the literature, such as [25]. However, we did not find such a strong association between arousal (or AU presence). This can be attributed to the MAE not being sufficiently small for arousal compared that for valence and this study using only limited types of faces. However, investigating the relationship between AUs and the valence-arousal dimension is beyond the scope of this research.

VI. DISCUSSION

We discuss several remaining issues as below:

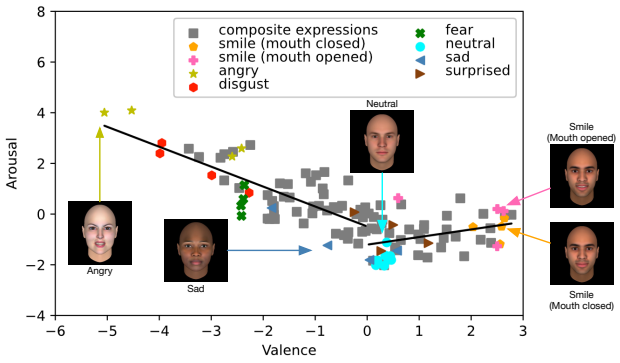

Fig. 3. Item parameter β estimated by the proposed model. A positive value means that it is difficult for the emotion image to obtain greater ratings; the directions of both axes are reversed for readability.

We used computer-generated faces as items. However, the proposed framework can be mathematically applied to other types of still images. For example, we used the VGG pre-trained on the in-the-wild facial expression dataset FER2013 [14], and therefore, our model arguably has the potential to handle in-the-wild facial expressions when sufficient data are available on the main task for fine-tuning. We believe that the number of items (i.e., 120 images) was sufficient as the sample space because a good test performance was obtained. The number of respondents meets the suggestion in [17], which concluded that 50 to 100 respondents would suffice for the same idiosyncratic perception prediction task.

Another issue is solving the technical limitations of the CORAL framework to integrate with less restricted IR models. There are several reasons for choosing the 1P-RS-GRM as the base IR model. 1P-RS-GRM is a cumulative logit IR model, which is inherently ordinal and thus more psychologically interpretable than the adjacent-category (or local logit) ones; cumulative logit imposes an inequality constraint on parameters where cut-off values for determining the rank on perceptual representation are monotonically increasing and thus reasonably interpretable [46]. However, 1P-RS-GRM is the most constrained cumulative logit model, and there exist less constrained models such as GRM. The linear component of the sigmoid function is expressed as $\alpha_i(\theta_j - \beta_{is})$ in the psychological literature [45]. Unfortunately, ensuring that the item-by-score parameter $\beta_{is} = f(\mathbf{x}_i)$ is rank consistent for all items in the CORAL framework complicates parameter optimization. Another candidate, RS-GRM, lies between the GRM and the 1P-RS-GRM. The linear component of RS-GRM is represented as $\alpha_i(\theta_j - \beta_i - \kappa_s)$, where the slope parameter α_i (positive value) indicates how well the item discriminates the abilities and sensitivities of the respondents. Extending our model to these IR models is an interesting direction for this study.

The IR theory is not the only choice to provide explanations of the model. For instance, in crowd aggregation or truth inference, wherein the goal is to obtain the ground truth from labels given by a crowd of people more accurately than simple averaging or majority voting, it is common practice to employ both person (e.g., skill/expertise, bias) and item parameters (e.g., difficulty) [55]. Some crowd aggregation models are based on psychological theories such as the signal detection theory [48]. However, crowd aggregation techniques are used primarily as a preprocessing step before building a main (deep) prediction model to obtain cleaner and more accurate ground truth labels without the use of any auxiliary information [44]. Therefore, they are closely related to the descriptive (nonexplanatory) IR theory.

Although we used the VGG as a latent regressor, various other FER models can be used as a replacement. The CNN was sufficient for the preliminary evaluation of the proposed framework using discrete ratings for both time and emotional dimensions. However, further evaluation and model extensions are required to investigate whether our framework can be applied to continuous annotations, as is the recent trend in the AC community, for example, [10], [52]. For the time domain, the CNN can be replaced by various video-based models using recurrent neural networks such as the long short-term memory and gated recurrent unit. As for space-continuous ratings, other types of IR models such as the continuous response model [36] would be applicable.

Though the performance of the most interpretable model (AU-RS-GRM) was reasonably lower than that of the proposed model, it is still an option for researchers who seek an easy-to-use, highly interpretable model that does not require time-consuming fine-tuning. However, the advantages of our approach can be further enhanced when it is integrated with post-hoc approaches such as GradCAM [41] to provide some explanations about the latent regression layer, i.e., CNN, while retaining the overall prediction performance.

Finally, we regressed only the item parameter from the images. In explanatory IR theory, it is also possible to latent-regress the person parameter from the demographic of the respondent (e.g., gender and ethnicity) and their personality traits [49]. Several limited efforts have been made to incorporate such personal information into the prediction of the emotional perception of an individual. For example, in [17], [19], the rating tendencies of the respondents, which can be considered a respondent parameter, were linked with their gender and personality trait scores using a probabilistic topic model. However, building deep neural networks for this purpose is not easy because of the lack of publicly available pre-trained models.

VII. Conclusion

We proposed CORAL-RS-GRM, which achieved high prediction performance and psychological interpretability in the problem of idiosyncratic perceived emotion recognition. The proposed model parameters in the IR layer were as interpretable as those of the corresponding base IR model, and the CNN realized reliable automatic feature extraction in the latent regression layer. The intrinsic approach can be employed be

a complementary framework for existing post-hoc approaches toward XAC to coach/support human social interactions.

REFERENCES

[1] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[2] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: facial behavior analysis toolkit," in *FG*. IEEE, 2018, pp. 59–66.

[3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, and T. et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai," *arXiv*, pp. arXiv–1910, 2019.

[4] F. Bartolucci, S. Bacci, and M. Gnaldi, *Statistical analysis of questionnaires: A unified approach based on R and Stata*. CRC Press, 2015, vol. 34.

[5] R. A. Calvo and D. Peters, *Positive computing: technology for wellbeing and human potential*. MIT Press, 2014.

[6] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognit. Lett.*, vol. 140, pp. 325–331, 2020.

[7] S. Cheng, Q. Liu, E. Chen, Z. Huang, Z. Huang, Y. Chen, H. Ma, and G. Hu, "DIRT: Deep learning enhanced item response theory for cognitive diagnosis," in *Proc. ACM CIKM*, 2019, pp. 2397–2400.

[8] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *Proc. CVPR*. IEEE, 2013, pp. 3515–3522.

[9] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, no. 1-2, pp. 5–32, 2003.

[10] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA Tut. Res. Wksh. Spch. Emo.*, 2000.

[11] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing emotion: an overview," *IJSE*, vol. 3, no. 1, pp. 1–17, 2012.

[12] V. E. Farrugia, H. P. Martínez, and G. N. Yannakakis, "The preference learning toolbox," 2015.

[13] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *P. A.*, vol. 3, no. 2, p. 5, 1978.

[14] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, and B. e. a. Hamner, "Challenges in representation learning: A report on three machine learning contests," in *Proc. NIPS*, 2013, pp. 117–124.

[15] G. Greene, "The ethics of ai and emotional intelligence," *Partnership on AI*, 2020.

[16] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vision Comput.*, vol. 31, no. 2, pp. 120–136, 2013.

[17] S. Kumano, R. Ishii, and K. Otsuka, "Computational model of idiosyncratic perception of others' emotions," in *Proc. ACII*, 2017, pp. 42–49.

[18] S. Kumano and K. Nomura, "Multitask item response models for response bias removal from affective ratings," in *ACII*, 2019, pp. 1–7.

[19] S. Kumano, K. Otsuka, M. Matsuda, R. Ishii, and J. Yamato, "Using a probabilistic topic model to link observers' perception tendency to personality," in *Proc. ACII*, 2013, pp. 588–593.

[20] P. Kuppens, F. Tuerlinckx, J. A. Russell, and L. F. Barrett, "The relation between valence and arousal in subjective experience." *Psychol. Bull.*, vol. 139, no. 4, p. 917, 2013.

[21] P. J. Lang, M. M. Bradley, B. N. Cuthbert *et al.*, "International affective picture system (iaps): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, vol. 1, pp. 39–58, 1997.

[22] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Compup.*, 2020.

[23] S. Liu, Q. Tan, S. Han, W. Li, X. Wang, Y. Gan, Q. Xu, X. Zhang, and L. Zhang, "The language context effect in facial expressions processing and its mandatory characteristic," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[24] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proc. CVPRW*, 2010, pp. 94–101.

[25] M. Mehu and K. R. Scherer, "Emotion categories and dimensions in the facial communication of affect: An integrated approach." *Emotion*, vol. 15, no. 6, p. 798, 2015.

[26] D. Melhart, K. Sfikas, G. Giannakakis, and G. Y. A. Liapis, "A study on affect model validity: Nominal vs ordinal labels," in *Wksh. Artif. Intell. Affect. Comput.* PMLR, 2020, pp. 27–34.

[27] 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, 2017.

[28] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *Proc. CVPR*. IEEE, 2016, pp. 4920–4928.

[29] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. ICME*. IEEE, 2005, pp. 5–pp.

[30] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning." in *Proc. Interspeech*, vol. 2017, 2017, pp. 1103–1107.

[31] R. W. Picard, *Affective computing*. MIT press, 2000.

[32] R. G. Praveen, E. Granger, and P. Cardinal, "Deep domain adaptation with ordinal regression for pain assessment using weakly-labeled videos," *Image Vis. Comput.*, p. 104167, 2021.

[33] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[34] J. A. Russell, "A circumplex model of affect." *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–178, 1980.

[35] J. A. Russell and J. M. F. Dols, *The psychology of facial expression*. CUP, 1997, vol. 10.

[36] F. Samejima, "Estimation of latent ability using a response pattern of graded scores." *Psychometrika monograph supplement*, 1969.

[37] S. Savvidou, "Validation of the feeltrace tool for recording impressions of expressed emotion," Ph.D. dissertation, Queen's Univ. Belfast, 2011.

[38] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, no. 1-2, pp. 227–256, 2003.

[39] K. R. Scherer and G. Ceschi, "Lost luggage: a field study of emotion–antecedent appraisal," *Motiv. Emot.*, vol. 21, no. 3, pp. 211–235, 1997.

[40] B. Schuller, "Responding to uncertainty in emotion recognition," *J. Inf. Commun. Ethics Soc.*, 2019.

[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. ICCV*, 2017, pp. 618–626.

[42] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," *arXiv preprint arXiv:1909.00360*, 2019.

[43] W. Shen, B. Wang, Y. Jiang, Y. Wang, and A. Yuille, "Multi-stage multi-recursive-input fully convolutional networks for neuronal boundary detection," in *Proc. ICCV*, Oct 2017.

[44] V. S. Sheng and J. Zhang, "Machine learning with crowdsourcing: A brief summary of the past research and future directions," in *Proc. AAAI Conf. AI*, vol. 33, no. 01, 2019, pp. 9837–9843.

[45] A. Suzuki, T. Hoshino, and K. Shigemasu, "Measuring individual differences in sensitivities to basic emotions in faces," *Cognition*, vol. 99, no. 3, pp. 327–353, 2006.

[46] F. Tuerlinckx and W.-C. Wang, "Models for polytomous data," in *Explanatory Item Response Models*. Springer, 2004.

[47] M. Uto and Y. Uchida, "Automated short-answer grading using deep neural networks and item response theory," in *AIED*. Springer, 2020, pp. 334–339.

[48] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, "The multidimensional wisdom of crowds," in *Proc. NIPS*, 2010, pp. 2424–2432.

[49] M. Wilson, P. De Boeck, and C. H. Carstensen, "Explanatory item response models: A brief introduction," in *Assessment of Competencies in Educational Contexts*. Hogefe & Huber Göttingen, Germany, 2008.

[50] X. Y, "Emotion_classification: facial expression recognition with pytorch," *GitHub repository*, 2020.

[51] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Trans. Affect. Comput.*, vol. 12, no. 01, pp. 16–35, jan 2021.

[52] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *Proc. ACII*, 2015, pp. 574–580.

[53] C.-K. Yeung, "Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory," *arXiv preprint arXiv:1904.11738*, 2019.

[54] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, "Learning personalized models for facial expression analysis and gesture recognition," *Trans. Multimed.*, vol. 18, no. 4, pp. 775–788, 2016.

[55] J. Zhang, X. Wu, and V. S. Sheng, "Learning from crowdsourced labeled data: a survey," *Artif. Intell. Rev.*, vol. 46, no. 4, pp. 543–576, 2016.
[56] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *Proc. ICASSP*, 2017, pp. 4990–4994.
[57] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," *TOMM*, vol. 15, no. 3s, pp. 1–32, 2019.
[58] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, "Predicting personalized image emotion perceptions in social networks," *Trans. ACII*, vol. 9, no. 4, pp. 526–540, 2016.