Inconsistency-Based Multi-Task Cooperative Learning for Emotion Recognition

Yifan Xu, Yuqi Cui, Xue Jiang, Yingjie Yin, Jingting Ding, Liang Li, and Dongrui Wu, *Senior Member, IEEE*

Abstract—Emotion recognition is an important part of affective computing. Human emotions can be described categorically or dimensionally. Accurate machine learning models for emotion classification and estimation usually depend on a large amount of annotated data. However, label acquisition in emotion recognition is costly: obtaining the ground-truth labels of an emotional sample usually requires multiple annotators' assessments, which is expensive and time-consuming. To reduce the labeling effort in multi-task emotions recognition, the paper proposes an inconsistency measure that can indicate the difference between the labels estimated from the feature space and the label distribution of labeled dataset. Using the inconsistency as an indicator of sample informativeness, we further propose an inconsistency-based multi-task cooperative learning framework that integrates multi-task active learning and self-training semi-supervised learning. Experiments in two multi-task emotion recognition scenarios, multi-dimensional emotion estimation and simultaneous emotion classification and estimation, were conducted under this framework. The results demonstrated that the proposed multi-task active learning framework outperformed several single-task and multi-task active learning approaches.

Index Terms—Active learning, semi-supervised learning, multi-task learning, cooperative learning, emotion recognition

1 Introduction

Affective computing aims to endow machines the ability to recognize, interpret, and synthesize human affects for harmonious human-machine interaction [1]. Emotion recognition is an important part of affective computing. It attempts to infer human emotions from various forms of inputs, e.g., facial expressions [2], gestures [3], speech [4], or physiological signals [5].

Human emotions can be described both categorically and dimensionally. Compared with intuitive categorical representations, such as Ekman's six basic emotions [6], dimensional representations are more suitable for characterizing continuous and fine-grained emotions. Commonly used dimensional emotion spaces include the Valence-Arousal 2D space [7] and the Valence-Arousal-Dominance 3D space [8]. This paper considers two multi-task emotion recognition scenarios: multi-dimensional emotion estimation (MDEE), and simultaneous emotion classification and estimation (SECE). In MDEE, we consider emotion estimation in the 3D Valance-Arousal-Dominance space. SECE considers further emotion classification, in addition to dimensional emotion estimation.

Collecting unlabeled affective samples is usually easy (e.g., videos and speeches can be easily recorded), but acquiring their labels is costly and time-consuming, due to the ambiguity and subjectiveness of emotions. Labeling affective samples in multi-task emotion recognition, where labels in different tasks need to be determined simultaneously, is particularly challenging. Active learning (AL) [9], [10] and semi-supervised learning (SSL) [11] are commonly used remedies.

AL uses different strategies to estimate the usefulness of unlabeled samples and selects the best ones to query for their labels; thus, better learning performance can be achieved from a small number of labeled samples. It has been used in both emotion classification and regression. Muhammad and Alhamid [12] selected samples with large entropy (low classification confidence) for labeling in facial emotion classification. Zhang et al. [13] selected samples with medium uncertainty in support vector machine for labeling in speech emotion classification, and dynamically allocated annotators for each sample until the user-specified annotation agreement level was met. Han et al. [14] transformed single dimensional emotion regression into a positive-negative binary classification problem, and selected the samples with high uncertainty in the classification model to annotate their dimensional labels. This approach helps improve the correlation coefficient of the regression model. Abdelwahab and Busso [15] evaluated the performance of an uncertainty-based AL algorithm and three greedy sampling-based ones (GSx, GSy and iGS in [16]) in valence and arousal estimation using deep neural networks, and demonstrated that greedy sampling in the feature space (GSx) can achieve both higher concordance correlation coefficient and lower variance. Wu and Huang [10] extended two greedy sampling-based

- *Yifan Xu, Yuqi Cui, Xue Jiang, and Dongrui Wu are with the Key Laboratory of the Ministry of Education for Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China. E-mail: {yfxu, yqcui, xuejiang, drwu}@hust.edu.cn.*
- *Yingjie Yin and Liang Li are with the Ant Group, World Financial Center, Beijing 100024, China. E-mail: {gaoshi.yyj, double.ll}@antgroup.com.*
- *Jingting Ding is with the Ant Group, Hangzhou 310023, China. E-mail: yimou.djt@antgroup.com.*

Manuscript received 18 March 2022; revised 9 July 2022; accepted 5 August 2022. Date of publication 9 August 2022; date of current version 15 November 2022. This research was supported in part by CCF-AFSG Research Fund under Grant RF20210007 and in part by Technology Innovation Project of Hubei Province of China under Grant 2019AEA171. (Corresponding author: Dongrui Wu.) Recommended for acceptance by C. Busso. Digital Object Identifier no. 10.1109/TAFFC.2022.3197414

single-task AL algorithms (GSy and iGS in [16]) to multi-task AL in MDEE, by considering the diversity of the three affect primitives simultaneously. It has been verified [10] that iGS is more efficient than expected model change maximization [17] and query-by-committee (QBC) [18] in AL for regression, and the multi-task version [16] further improve its performance in MDEE. Jiang et al. [19] used rank combination [20] that weights the AL ranks of all single tasks to select the most beneficial samples to label in SECE.

AL only selects a small number of samples to query for their labels. The remaining large amount of unlabeled samples in the data pool also contain rich information, which can be exploited through SSL. For example, self-training SSL first uses the model trained in the previous iteration to temporally label the samples, and then identifies those with high confidence and assigns pseudo-labels to them. Zhang et al. [21] proposed a cooperative learning approach that combines self-training and AL in speech emotion recognition. Experiments on two binary classification datasets verified the effectiveness of cooperative learning and its multi-view and mixed-view variants.

However, in dimensional emotion regression, it is difficult to directly compute the confidence of the outputs in the regression model and employ self-training like in emotion classification. This paper proposes an inconsistency measure that can indicate the difference between the labels estimated from the feature space and the conditional label distribution of the labeled samples, which only depends on the relationship between the label spaces of different tasks. The inconsistency can be viewed as an informativeness indicator: samples with large inconsistency can increase the label diversity of the labeled dataset.

Consider MDEE first. Given an emotional sample that is predicted to have low valence, high arousal, and low dominance, e.g., a sample with *fear* emotion, we can calculate its label inconsistency in the Dominance dimension using the labeled dataset, based on its estimated labels in the other two dimensions. First, we identify the labeled samples that have similar Valence and Arousal values with this sample and check their Dominance labels. Assume most of these samples have high Dominance, e.g., with *anger* emotion. Then, the given sample with low Dominance is inconsistent with the label distribution of these similar samples, and can thus increase the label diversity in Dominance. Similarly, we can obtain its label inconsistency with the labeled dataset in Valence and Arousal. Aggregating the label inconsistency in all three dimensions, we obtain the sample's total inconsistency with the labeled dataset.

For SECE, we measure the label inconsistency differently, since additional categorical labels are available. More specifically, the dimensional label distributions are estimated from the categorical labels, unlike in MDEE where the conditional label distribution of each dimension is estimated from the remaining two dimensions. For example, consider a sample which is estimated to have *surprise* emotion, i.e., high Valence, medium Arousal and medium Dominance. Assume that most of the labeled samples having the same categorical label with the given sample have high Valence, high Arousal and medium Dominance. Then, the given sample is inconsistent with the category-conditional label distribution in Arousal but consistent in Valence and Dominance.

Based on this informativeness measure, we further propose an inconsistency-based multi-task cooperative learning (IMCL) framework that integrates AL and SSL. Specifically, IMCL first computes the inconsistency of the unlabeled samples in the tasks where conditional label distribution can be estimated from other tasks, and integrates them into the total inconsistency. Then, it selects the most inconsistent sample to query for its label (i.e., it uses AL to select the most informativeness sample to manually label), and assigns pseudo-labels to samples with low inconsistency (i.e., it uses self-training SSL to label the high-confident samples, which are consistent with the current label distribution), to enlarge the labeled training set. The samples with manual annotations and pseudo-labels are subsequently combined and utilized to update corresponding task models. The overall flowchart of IMCL in a two-task dimensional emotion estimation application is illustrated in Fig. 1.

The contributions of this paper are:

1) We propose an informativeness measure to represent the inconsistency between the estimated labels of unlabeled samples and the true label distribution of labeled samples.
2) Based on the inconsistency measure, we further propose IMCL, a multi-task cooperative learning framework that integrates AL and SSL.
3) Experiments on two speech datasets and one image dataset verified that our proposed IMCL can effectively select valuable samples for annotation and utilize unlabeled samples.

The remainder of the paper is organized as follows: Section 2 introduces the framework of our proposed IMCL approach. Section 3 describes the datasets and implementation details of IMCL in MDEE and SECE in the experiments. Section 4 compares the performance of IMCL with other AL approaches in MDEE and SECE, and discusses the results. Section 5 draws conclusions and points out some future research directions.

2 Inconsistency-Based Multi-Task Cooperative Learning (IMCL)

This section introduces our proposed IMCL framework.

2.1 Problem Setting

Assume the Task Set $\mathcal{T}$ contains more than one tasks in emotion recognition. In MDEE, $\mathcal{T} = \{$Valence estimation, Arousal estimation, Dominance estimation$\}$, and in SECE, $\mathcal{T} = \{$ Emotion classification, Valence estimation, Arousal estimation, Dominance estimation$\}$. The data pool consists of a small number of labeled samples $X^L = \{(\boldsymbol{x}_i^L, \boldsymbol{y}_{i,\mathcal{T}}^L)\}_{i=1}^{N_L}$ and a large number of unlabeled samples $X^U = \{\boldsymbol{x}_j^U\}_{j=1}^{N_U}$, where $\boldsymbol{x}_i \in \mathbb{R}^{d\times 1}$ is a d-dimensional feature vector, and $\boldsymbol{y}_i \in \mathbb{R}^{|\mathcal{T}|\times 1}$ its $|\mathcal{T}|$-dimensional label vector corresponding to the tasks in $\mathcal{T}$.

2.2 Emotion Recognition Model f_t

For each task $t \in \mathcal{T}$, we train a ridge regression (RR) model for emotion regression, or a logistic regression (LR) model for emotion classification, $f_t(\boldsymbol{x})$, on $\{(\boldsymbol{x}_i^L, y_{i,t}^L)\}_{i=1}^{N_L}$, using the original features as inputs to estimate the labels of the unlabeled samples.

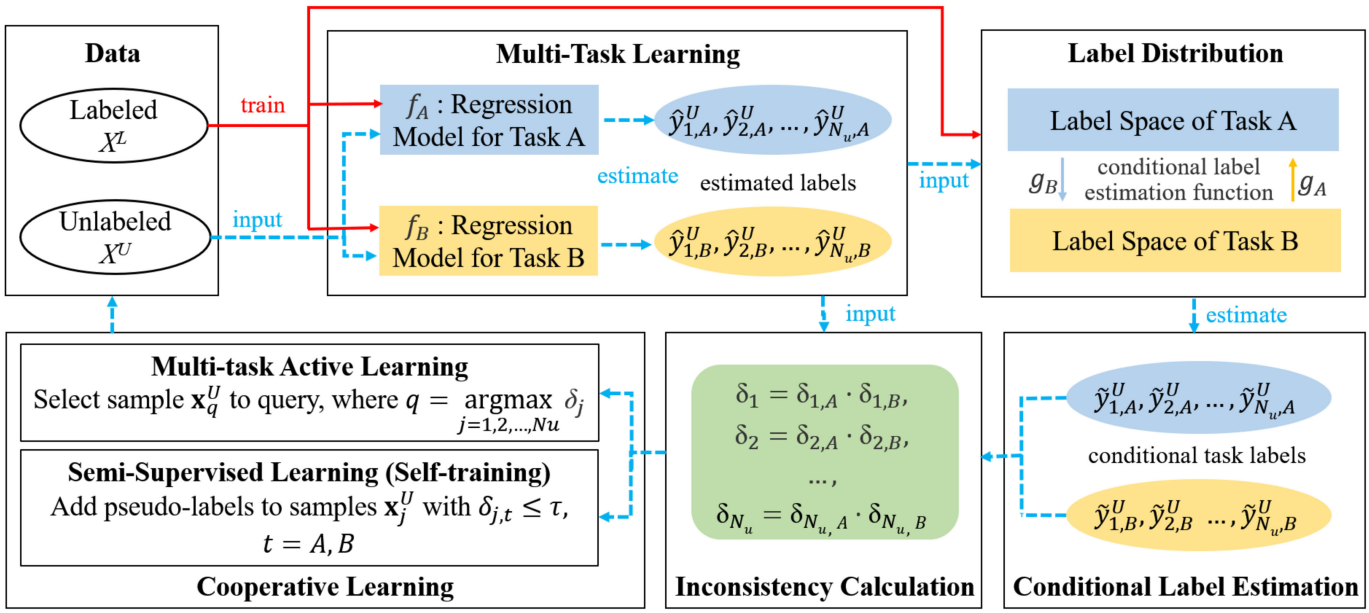

Fig. 1. The flowchart of IMCL in two-task dimensional emotion estimation. It first trains regression models f_A and f_B for Tasks A and B, separately, using the labeled samples, and uses them to estimate labels of the unlabeled samples. Then, it uses the labeled samples to construct the conditional label estimation function g_A and g_B between the label spaces of the two tasks. The conditional task labels of the unlabeled samples are obtained through g_A and g_B from their estimated labels. With estimated labels and conditional task labels, the inconsistency of each unlabeled sample can be computed. IMCL next selects the most inconsistent sample to query for its groundtruth label, and assigns pseudo-labels to some other samples with low inconsistency. These selected samples with true labels are then added to the training set and combined with the pseudo-labeled samples to update the regression models f_A and f_B. Conditional label estimation functions g_A and g_B are next updated using manually labeled samples.

For an unlabeled sample $\boldsymbol{x}_j^U$, its estimated label vector is

$$\hat{\boldsymbol{y}}_{j,\mathcal{T}}^U = \left[\hat{y}_{j,1}^U; \ldots; \hat{y}_{j,|\mathcal{T}|}^U\right] = \left[f_1(\boldsymbol{x}_j^U); \ldots; f_{|\mathcal{T}|}(\boldsymbol{x}_j^U)\right]. \quad (1)$$

2.3 Conditional Label Distribution Model g_t

To represent the label distribution of the labeled dataset in Tasks $\mathcal{T}^{dis}$, we also construct a set of models $\{g_t\}_{t\in\mathcal{T}^{dis}}$. For a Task t in $\mathcal{T}^{dis}$, g_t takes the labels of the related tasks $\mathcal{T}_t^{rel}$ as inputs and outputs the conditional label distribution in Task t.

In MDEE, we estimate the conditional label distribution of each task from the remaining ones, i.e., $\mathcal{T}^{dis} = \mathcal{T} = \{$Valence estimation, Arousal estimation, Dominance estimation $\}$ and $\mathcal{T}_t^{rel} = \{\mathcal{T}\backslash t\}$. g_t adopts k-nearest neighbors (k NN) regressor with $k = 5$, which takes the estimated labels of an unlabeled sample $\boldsymbol{x}_j^U$ in $\mathcal{T}_t^{rel}$, i.e., $\hat{\boldsymbol{y}}_{j,\mathcal{T}_t^{rel}}^U$, as inputs, finds its nearest labeled neighbors in the label spaces of $\mathcal{T}_t^{rel}$, and averages their labels in Task t as the conditional task label $\tilde{y}_{j,t}^U$, i.e.,

$$\tilde{y}_{j,t}^U = g_t\left(\hat{\boldsymbol{y}}_{j,\mathcal{T}_t^{rel}}^U\right) = kNN\left(\hat{\boldsymbol{y}}_{j,\mathcal{T}_t^{rel}}^U\right), \quad t \in \mathcal{T}^{dis}. \quad (2)$$

In SECE, the conditional distributions of the dimensional emotions are computed from the estimated categorical emotion probabilities $\hat{\boldsymbol{y}}_C$, i.e., $\mathcal{T}^{dis} = \{$Valence estimation, Arousal estimation, Dominance estimation $\}$, and $\mathcal{T}_t^{rel} = \{$ Emotion classification$\}$. We obtain the conditional task labels of the dimensional emotions from the estimated categorical emotion probabilities $\hat{\boldsymbol{y}}_C$ through $g(\hat{\boldsymbol{y}}_C)$. Each element $\hat{y}_e$ in $\hat{\boldsymbol{y}}_C \in \mathbb{R}^{|E|\times 1}$ denotes the estimated probability for Emotion e in the emotion category set E (an example is $E = \{$angry, happy, excited, sad, frustrated$\}$, as in our experiments in the next section). $g(\hat{\boldsymbol{y}}_C)$ computes the average dimensional emotion values for each emotion category on the labeled samples, and multiplies them with the corresponding emotion classification probabilities in $\hat{\boldsymbol{y}}_C^U$.

Specifically, for an emotion category e in SECE (e.g., happy), let $y_{i,v}^L$, $y_{i,a}^L$ and $y_{i,d}^L$ be the valence, arousal and dominance values of $\boldsymbol{x}_i^L$, respectively; then, the average dimensional values are

$$\boldsymbol{h}_e = \frac{1}{\sum_{i=1}^{N^L} y_{i,e}} \sum_{i=1}^{N^L} y_{i,e}^L \cdot \left[y_{i,v}^L; y_{i,a}^L; y_{i,d}^L\right]. \quad (3)$$

$y_{i,e}^L \in \{0, 1\}$ denotes if $\boldsymbol{x}_i^L$ has Emotion e. $y_{i,e}^L = 1$ means $\boldsymbol{x}_i^L$ belongs to emotion category e and should be included in the calculation of the conditional labels for e, and vice versa.

The conditional task labels of the dimensional emotions for an unlabeled sample $\boldsymbol{x}_j^U$ are calculated by

$$\tilde{\boldsymbol{y}}_{j,\mathcal{T}^{dis}}^U = g(\hat{\boldsymbol{y}}_{j,C}^U) = \sum_{e\in E} \hat{y}_{j,e}^U \cdot \boldsymbol{h}_e. \quad (4)$$

Note that the original features are not considered at all in g_t.

2.4 Inconsistency Calculation

With the estimated label $\hat{y}_{j,t}^U$ from f_t and the conditional task label $\tilde{y}_{j,t}^U$ from g_t of Task $t \in \mathcal{T}^{dis}$, the inconsistency δ_j of an unlabeled sample $\boldsymbol{x}_j^U$ can then be computed. We employ two different inconsistency calculation functions for MDEE and SECE to demonstrate their flexibility.

In MDEE, the inconsistency δ_j is computed as

$$\delta_j = \sum_{t\in\mathcal{T}^{dis}} \delta_{j,t} = \sum_{t\in\mathcal{T}^{dis}} \left|\hat{y}_{j,t}^U - \tilde{y}_{j,t}^U\right|, \quad j = 1, \ldots, N_U. \quad (5)$$

In SECE, the inconsistency δ_j is computed as

$$\delta_j = \left\| \hat{\boldsymbol{y}}^U_{j,\mathcal{T}^{dis}} - \tilde{\boldsymbol{y}}^U_{j,\mathcal{T}^{dis}} \right\|_2, \quad j = 1, \ldots, N_U. \tag{6}$$

2.5 Inconsistency-Based Multi-Task AL (IMAL)

The inconsistency indicates the label difference of $t \in \mathcal{T}^{dis}$ between an unlabeled sample and the current labeled dataset X^L, i.e., the diversity a sample could bring to X^L. Thus, it can be used as an informativeness indicator to guide AL.

We propose inconsistency-based multi-task active learning (IMAL), which selects the unlabeled sample $\boldsymbol{x}^U_q$ with the maximum inconsistency and queries for its groundtruth label in all tasks, i.e.,

$$q = \arg\max_{j=1,\ldots N_U} \delta_j. \tag{7}$$

The pseudo-code of IMAL is shown in Algorithm 1. Both $\{f_t\}_{t\in\mathcal{T}}$ and $\{g_t\}_{t\in\mathcal{T}^{dis}}$ are updated iteratively during training and used to select samples for AL. In the test stage, only f_t is used.

Algorithm 1. Inconsistency-Based Multi-Task Active Learning (IMAL)

Input: Labeled training data $X^L = \{\boldsymbol{x}^L_i, \boldsymbol{y}^L_{i,\mathcal{T}}\}_{i=1}^{N_L}$;
Unlabeled training data $X^U = \{\boldsymbol{x}^U_j\}_{j=1}^{N_U}$;
K, number of samples to be queried.
Output: $|\mathcal{T}|$ emotion recognition models $\{f_t\}_{t\in\mathcal{T}}$.
for $t \in \mathcal{T}$ **do**
  Use $\{\boldsymbol{x}^L_i, y^L_{i,t}\}_{i=1}^{N_L}$ to train f_t;
end
for $k = 1 : K$ **do**
  Estimate $\{\hat{\boldsymbol{y}}^U_{j,\mathcal{T}}\}_{j=1}^{N_U}$ of X^U using (1);
  for $t \in \mathcal{T}^{dis}$ **do**
    Construct the conditional label estimation function g_t using $\{\boldsymbol{y}^L_{i,\mathcal{T}^{rel}_t}, y^L_{i,t}\}_{i=1}^{N_L}$;
    Obtain the conditional task labels $\{\tilde{y}^U_{j,t}\}_{j=1}^{N_U}$ of X^U using (2) for MDEE or (4) for SECE;
  end
  Compute the inconsistency $\{\delta_j\}_{j=1}^{N_U}$ using (5) for MDEE or (6) for SECE;
  Select the most inconsistent sample $\boldsymbol{x}^U_q$ using (7);
  Query for $\boldsymbol{y}^U_{q,\mathcal{T}}$, groundtruth labels of $\boldsymbol{x}^U_q$ in all the tasks;
  $X^U \leftarrow X^U \backslash \boldsymbol{x}^U_q, \quad N_U \leftarrow N_U - 1$;
  $X^L \leftarrow X^L \cup (\boldsymbol{x}^U_q, \boldsymbol{y}^U_{q,\mathcal{T}}), \quad N_L \leftarrow N_L + 1$;
  for $t \in \mathcal{T}$ **do**
    Use $\{\boldsymbol{x}^L_i, y^L_{i,t}\}_{i=1}^{N_L}$ to update f_t;
  end
end

2.6 Inconsistency-Based SSL

We adopt self-training SSL to exploit the unlabeled samples with high consistency between the estimated labels and the conditional task labels. These samples are selected and assigned pseudo-labels. Specifically, a subset of samples selected under some user-specified rules, e.g., samples with top-p% minimum inconsistency, are automatically assigned pseudo-labels, and then added to the sample set X^P_t and incorporated into the training set to train the corresponding task model f_t.

The pseudo-label of $\boldsymbol{x}^U_j$ in Task $t \in \mathcal{T}^{dis}$ is calculated by

$$\bar{y}_{j,t} = \alpha \times \hat{y}^U_{j,t} + (1 - \alpha) \times \tilde{y}^U_{j,t}, \tag{8}$$

where $\alpha \in [0, 1]$ is the weight of the estimated labels in the pseudo-labels, which is a hyper-parameter to be specified by the user. For a large α, the estimated labels from f_t would dominate the pseudo-labels. On the contrary, a small α enables the conditional task labels from g_t to dominate the pseudo-labels.

2.7 IMCL

IMCL integrates IMAL and SSL to utilize both manually and pseudo- labeled samples. Its pseudo-code is given in Algorithm 2. We attempted to adopt different inconsistency measures and self-training sample selection rules in MDEE and SECE to verify the flexibility of IMCL. More implementation details of IMCL in MDEE and SECE can be found in Section 3.3.

Algorithm 2. Inconsistency-Based Multi-Task Cooperative Learning (IMCL)

Input: Labeled training data $X^L = \{\boldsymbol{x}^L_i, \boldsymbol{y}^L_{i,\mathcal{T}}\}_{i=1}^{N_L}$;
Unlabeled training data $X^U = \{\boldsymbol{x}^U_j\}_{j=1}^{N_U}$;
α, weight of the estimated label in (8);
K, number of samples to be queried;
Sample selection rules in SSL.
Output: $|\mathcal{T}|$ emotion recognition models $\{f_t\}_{t\in\mathcal{T}}$.
for $t \in \mathcal{T}$ **do**
  Use $\{\boldsymbol{x}^L_i, y^L_{i,t}\}_{i=1}^{N_L}$ to train f_t;
end
for $k = 1 : K$ **do**
  Estimate $\{\hat{\boldsymbol{y}}^U_{j,\mathcal{T}}\}_{j=1}^{N_U}$ of X^U using (1);
  Initialize X^P_t to $\emptyset$, $t \in \mathcal{T}^{dis}$;
  for $t \in \mathcal{T}^{dis}$ **do**
    Construct the conditional label estimation function g_t using $\{\boldsymbol{y}^L_{i,\mathcal{T}^{rel}_t}, y^L_{i,t}\}_{i=1}^{N_L}$;
    Obtain the conditional task labels $\{\tilde{y}^U_{j,t}\}_{j=1}^{N_U}$ of X^U using (2) for MDEE or (4) for SECE;
  end
  for $j = 1 : N_U$ **do**
    for $t \in \mathcal{T}^{dis}$ **do**
      Add sample $\boldsymbol{x}^U_j$ and its corresponding pseudo-label $\bar{y}_{j,t}$ computed by (8) to X^P_t, if it satisfies sample selection rules in SSL;
    end
  end
  Compute the inconsistency $\{\delta_j\}_{j=1}^{N_U}$ using (5) for MDEE or (6) for SECE;
  Select the most inconsistent sample $\boldsymbol{x}^U_q$ using (7);
  Query for $\boldsymbol{y}^U_{q,\mathcal{T}}$, groundtruth labels of $\boldsymbol{x}^U_q$ in all the tasks;
  $X^U \leftarrow X^U \backslash \boldsymbol{x}^U_q, \quad N_U \leftarrow N_U - 1$;
  $X^L \leftarrow X^L \cup (\boldsymbol{x}^U_q, \boldsymbol{y}^U_{q,\mathcal{T}}), \quad N_L \leftarrow N_L + 1$;
  for $t \in \mathcal{T}$ **do**
    if $t \in \mathcal{T}^{dis}$ **do**
      Use $\{\boldsymbol{x}^L_i, y^L_{i,t}\}_{i=1}^{N_L} \cup X^P_t$ to update f_t;
    else
      Use $\{\boldsymbol{x}^L_i, y^L_{i,t}\}_{i=1}^{N_L}$ to update f_t;
    end
  end
end

TABLE 1
Characteristics of the Three Affective Computing Datasets

Dataset	Size	d	Valence (mean $\pm$ std)	Arousal (mean $\pm$ std)	Dominance (mean $\pm$ std)
VAM	947	46	-0.2282 $\pm$0.1991	0.0280 $\pm$0.3425	0.0924 $\pm$0.3025
IAPS	1,178	30	5.0314 $\pm$1.7708	4.8159 $\pm$1.1509	5.1580 $\pm$1.0811
IEMOCAP	2,815	35	2.8272 $\pm$1.0576	3.1829 $\pm$0.7606	3.2481 $\pm$0.8077

'Size' denotes the number of samples, and d is the feature dimensionality.

Both $\{f_t\}_{t\in\mathcal{T}}$ and $\{g_t\}_{t\in\mathcal{T}^{dis}}$ are updated iteratively during training and used to select samples for AL and SSL. In the test stage, only f_t is used.

3 Datasets and Experimental Setup

This section presents experimental results to demonstrate the effectiveness of the proposed IMAL and IMCL.

3.1 Datasets and Feature Extraction

We verified the performance of our proposed IMAL and IMCL in MDEE on three public affective computing datasets, VAM, IAPS and IEMOCAP, with Valence-Arousal-Dominance dimensional emotion annotations. Their characteristics are shown in Table 1. Experiments on SECE were conducted on the IEMOCAP dataset, which has both categorical and dimensional emotion annotations.

The VAM corpus [22] contains 947 emotional utterances collected from 47 guests (11m/36f) in a German TV talk-show *Vera am Mittag* (*Vera at Noon* in English). The weighted average values of emotion primitives from several evaluators were used as the ground-truth labels of each sentence. We used 46 acoustic features, including nine pitch features, five duration features, six energy features, and 26 Mel Frequency Cepstral Coefficient (MFCC) features, as in [4], [10], [23]. Each feature was normalized to mean 0 and standard deviation 1.

The IAPS (International Affective Picture System) dataset [24] used documentary-style natural color images that can evoke strong emotions as stimuli. It includes 1,182 images, but we removed four duplicate ones. The mean valence, arousal and dominance values collected from multiple annotators were used as ground-truth labels. ResNet50 [25] pre-trained on ImageNet [26] was used as the encoder to extract features from the images. Principal component analysis was used to reduce the dimensionality to 30. Each feature was then z-normalized.

IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) [27] is a multi-modal dataset annotated with both categorical and dimensional emotion labels. We used only the audio modality. Because IEMOCAP has both categorical and dimensional annotations, it was used in experiments of both MDEE and SECE. Only five emotion categories ('angry', 'happy', 'excited', 'sad' and 'frustrated' with 289, 284, 663, 608 and 971 samples, respectively) were selected. 35 features were used [19], including two signal amplitude features (mean and standard deviation), two energy features, one pause feature, one harmonics feature, two pitch features, one zero-crossing rate feature, and 26 MFCC features. Each feature was also z-normalized. The datasets used in MDEE and SECE were identical, except that MDEE only used the dimensional labels, whereas SECE further used the categorical emotion label.

3.2 Experimental Setup

We compared the performances of the following seven sample selection strategies:

1) Baseline-all (BL-all), which assumes all samples in the data pool are labeled, and uses them to construct the models for each task. This represents the performance upper bound that an algorithm can achieve.
2) Random sampling (Rand), which selects all K samples randomly to annotate.
3) Single-task AL for Task t (ST-t). In MDEE, iGS [16] was used for each single emotion regression task. iGS uses greedy sampling in both feature and label spaces. In SECE, only single-task AL for classification (denoted as ST) was considered. A classical uncertainty-based AL algorithm that selects the sample with the least maximum classification confidence for annotation was used.
4) Multi-task iGS (MT-iGS) [10] that extends the single-task iGS to multi-task learning, by considering feature diversity and label diversity in the three affect primitives simultaneously.
5) Rank combination (RankComb) [19], which computes the weighted average of each sample's AL ranks in different tasks and then selects the sample with the smallest overall rank to annotate. This approach was only used as a baseline in SECE. The AL approach in emotion classification was identical to that in ST, and in each emotion estimation task was iGS [16]. The parameter settings, i.e., the rank weights of the tasks, were identical to those in [19].
6) IMAL, which was introduced in Algorithm 1 that uses inconsistency to select the samples for human annotation.
7) IMCL, which was introduced in Algorithm 2. The weight α in calculating the pseudo-labels in (8) was set to 0.5.

3.3 Implementation Details

In MDEE, we used Ridge Regression (RR) as the base model in each dimensional emotion estimation task. The weight of the regularization term in RR was set to $10/N^L$, as in [10].

MDEE used thresholding to identify the samples with high consistency and assigned them pseudo-labels. For each Task t, the samples with inconsistency $\delta_{j,t}$ computed from (5) no larger than a user-specified threshold τ_t was assigned pseudo-labels using (8). We set the threshold τ_t in Task t to $e_t/2$, where e_t was the root mean squared error (RMSE) of f_t on X^L.

In SECE, we used a Logistic Regression base model for emotion classification, in addition to the three RR base models for Valence-Arousal-Dominance estimation. The weight of the regularization term in all models was set to $10/N^L$.

SECE used a different sample selection strategy: top-$p\%$ (p was set to 5 in the experiments) unlabeled samples with the minimum inconsistency calculated from (6) were assigned pseudo-labels in all three dimensional emotions. Here we set both p in SECE and τ_t in MDEE empirically to a median value. We did not assign pseudo-labels of categorical emotion to the unlabeled samples.

Note that different conditional label estimation functions in MDEE and SECE were designed according to the characteristics of the task scenario, and different choices of aggregation functions and self-training sample selection rules were used to verify the flexibility of IMAL and IMCL.

3.4 Performance Evaluation

For the two speech datasets, we consider the cross-speaker scenario, i.e., the speakers in the training set do not overlap at all with those in the test set.

In VAM, each speaker has different numbers of samples, ranging from 4 to 46. We first randomly shuffled the indices of the speakers, and then added all samples from each individual speaker according to the shuffled order one by one, until the total number of selected samples reached 30% of the original dataset size (947). Finally, we used these $30\%{\cdot}947$ samples as the training set and the remaining $70\%{\cdot}947$ samples for test to validate the performances of different algorithms. This process was repeated 100 times, and the average results were reported.

IEMOCAP has five sessions, each including emotional samples from two-speaker interactions. Each time we used two sessions as the training set and the remaining three as the test set, and then switched the training and test set, resulting in 10 different training-test partitions in total. Experiments on each dataset partition were repeated ten times with different initial labeled samples.

For IAPS, the image dataset, we randomly selected 30% of the samples as the training set, and the remaining 70% as the test set, and repeated this process 100 times.

For all three datasets, the initial $d+1$ labeled samples were selected randomly, where d is the feature dimensionality of each dataset, as in [9]. Then, we iteratively selected samples from the unlabeled data pool by different algorithms to annotate, and updated the emotion recognition models accordingly.

We used RMSE and correlation coefficient (CC) to evaluate the performance of the regression models, where RMSE directly indicates the prediction error and was our primary performance measure. In emotion classification, weighted accuracy (the average of the per-class accuracies) was used to measure the performance of the classification models, eliminating the influence of class imbalance. The average RMSEs and CCs in emotion estimation, and weighted accuracies in emotion classification, are reported. 200 iterations of sample selection were performed on VAM and IAPS, and 300 on IEMOCAP due to its larger size.

4 Results and Discussions

The section presents our experimental results in MDEE and SECE, parameter sensitivity analysis, and additional discussions.

4.1 Experimental Results in MDEE

The performances of different sample selection approaches in MDEE are shown in Fig. 2. To emphasize the performance differences among different algorithms at the initial sampling stage, the horizontal axis used logarithm scale.

Fig. 2 shows that:

1) As the number of labeled samples increased, the performance of all models improved and gradually converged to BL-all. This is intuitive, since more labeled samples result in more reliable regression models.
2) All three single-task AL algorithms performed much better than random sampling on all three tasks, no matter which task it focused on. For a particular Task t, ST-t always outperformed ST-t' ($t' \neq t$). The results verified the effectiveness of single-task iGS even in multi-task scenarios, due to the intrinsic relationship among the tasks. We can also observe that among the three single-task AL algorithms, whereas usually ST-t ranked first in Task t (the task it focused on), it ranked poorly in the other two tasks. This is due to the fact that ST-t emphasized too much on a single task, and hence may sacrifice its performances on other tasks, i.e., single-task AL cannot achieve good compromise among multiple tasks.
3) MT-iGS takes all three tasks into consideration simultaneously, instead of emphasizing only one of them. For a particular Task t, MT-iGS either achieved comparable performance with ST-t, or performed only slightly worse than ST-t but much better than ST-t' ($t' \neq t$). Its overall superior performance on all three tasks demonstrated the advantage of multi-task AL over single-task AL: it can achieve a better trade-off among multiple tasks.
4) Our proposed IMAL achieved comparable performance with MT-iGS on all three datasets, since the sample selection criteria of these two approaches are both based on sample diversity. This demonstrates that the proposed inconsistency is a valid measure of informativeness in AL. The average performance of IMAL on all three tasks was generally better than that of the single-task AL approaches.
5) IMCL always reduced the RMSE of IMAL, many times also increased the CC. However, the increase of CC was not guaranteed, because the training of the regression models explicitly minimized the RMSE, but no objective was imposed on the CC directly. Surprisingly, IMCL sometimes even outperformed the best-performing single-task AL, suggesting the benefits of SSL.

Tables 2, 3, and 4 respectively show the average RMSEs on all three emotion dimensions at the 5th, 10th, 20th and 50th iterations on the three datasets. The performance gaps among different algorithms became smaller when the number of labeled samples increased, so the results after more iterations are omitted. IMCL achieved the lowest RMSEs on all three datasets.

To check if the performance improvement of IMCL over other approaches were statistically significant, paired t-tests with Holm's p-value adjustment [28] were performed. Those with the adjusted p-values smaller than 0.05 are marked with

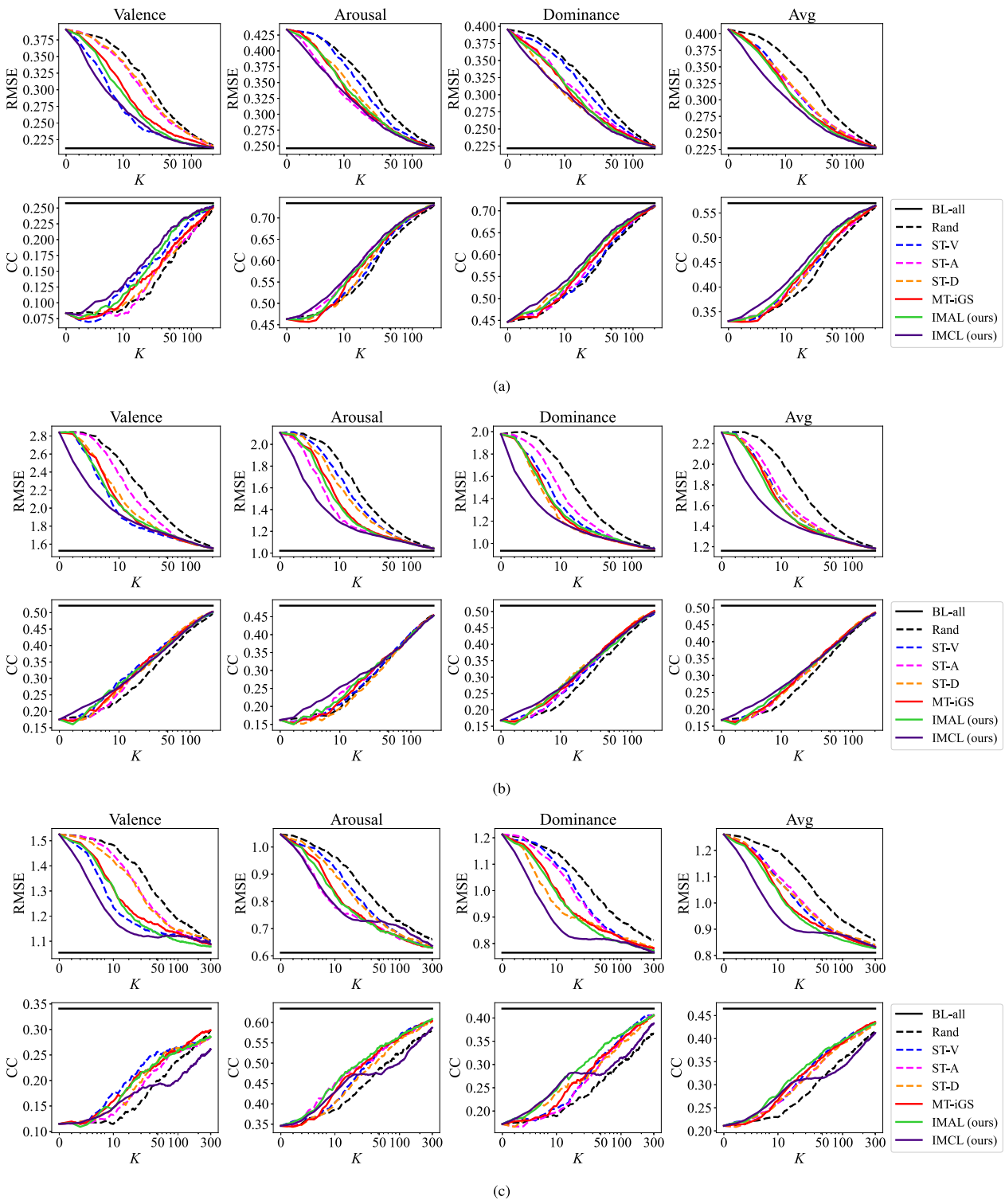

Fig. 2. Average performance of different sample selection algorithms in MDEE on (a) VAM, (b) IAPS, (c) IEMOCAP. Generally our proposed IMCL achieved the best performance, and IMAL the second best. K is the number of samples selected by different strategies to be manually annotated, in addition to the initial $d+1$ randomly selected labeled samples.

* in Tables 2, 3, and 4. Most results were statistically significant, especially at the initial sample selection stages, suggesting that IMCL was efficient in selecting valuable unlabeled samples for annotation and utilizing unsupervised information.

In summary, on average, IMCL performed the best among all algorithms, demonstrating the effectiveness of integrating AL and SSL.

4.2 Experimental Results in SECE

The average performances in SECE on the IEMOCAP dataset are shown in Fig. 3. Logarithm scale was also used on the horizontal axis. The classification performance of different approaches did not vary much.

As the dimensional emotions are more fine-grained and contain richer information, we mainly focus on the performances of the dimensional regression tasks. The average RMSEs

TABLE 2
Average RMSEs of the Three Emotion Dimensions at the 5th, 10th, 20th and 50th Iterations, on VAM in MDEE

Iteration	5	10	20	50
Rand	0.3876*	0.3654*	0.3352*	0.2796*
ST-V	0.3643*	0.3302*	0.2973*	0.2586*
ST-A	0.3602*	0.3300*	0.2994*	0.2610*
ST-D	0.3594*	0.3341*	0.3012*	0.2605*
MT-iGS	0.3559*	0.3206*	0.2855	0.2555*
IMAL (ours)	0.3532*	0.3186*	0.2858	0.2483
IMCL (ours)	0.3298	0.3041	0.2785	0.2484

** means IMCL outperformed the corresponding approach significantly in paired t-test with Holm's p-value adjustment ($\alpha = 0.05$).*

TABLE 4
Average RMSEs of the Three Emotion Dimensions at the 5th, 10th, 20th and 50th Iterations, on IEMOCAP in MDEE

Iteration	5	10	20	50
Rand	1.2232*	1.1975*	1.1302*	1.0083*
ST-V	1.1731*	1.0854*	1.0087*	0.9140*
ST-A	1.1610*	1.1031*	1.0307*	0.9287*
ST-D	1.1536*	1.0865*	1.0210*	0.9303*
MT-iGS	1.1579*	1.0512*	0.9664*	0.9072*
IMAL (ours)	1.1351*	1.0441*	0.9550*	0.8869
IMCL (ours)	1.0362	0.9470	0.8951	0.8860

** means IMCL outperformed the corresponding approach significantly in paired t-test with Holm's p-value adjustment ($\alpha = 0.05$).*

at the 5th, 10th, 20th and 50th iterations are shown in Table 5. Paired t-tests with Holm's p-value adjustment were also performed to check if IMCL significantly outperformed others. Those with the adjusted p-values smaller than 0.05 are marked with *.

The following observations can be made from Fig. 3 and Table 5 in SECE, similar to those in MDEE:

1) Different sample selection approaches achieved similar performance in emotion classification. Although single-task AL on emotion classification may not benefit the classification performance much, it still helped improve the emotion estimation performance, demonstrating that there exists some intrinsic relationship between categorical and dimensional emotions.
2) MT-iGS outperformed both RankComb and ST, indicating again that it is a very strong approach.
3) Our proposed IMAL outperformed MT-iGS slightly, but other AL approaches by a large margin. IMCL further improved IMAL by incorporating samples with pseudo-labels through SSL, achieving the best performance.

4.3 Parameter Sensitivity

There are two hyper-parameters in our proposed IMCL in Algorithm 2: α, weight of the estimated label in (8), and the sample selection rules in SSL, i.e., threshold vector $\boldsymbol{\tau}$ in MDEE and percentage p in SECE. Experiments in MDEE were carried out to analyze their influence.

TABLE 3
Average RMSEs of the Three Emotion Dimensions at the 5th, 10th, 20th and 50th Iterations, on IAPS in MDEE

Iteration	5	10	20	50
Rand	2.2146*	2.0333*	1.7496*	1.4337*
ST-V	1.9425*	1.6552*	1.4624*	1.3124*
ST-A	2.0154*	1.7179*	1.5082*	1.3229*
ST-D	1.9329*	1.6666*	1.4657*	1.3078*
MT-iGS	1.9038*	1.5956*	1.4071*	1.2928
IMAL (ours)	1.8558*	1.5917*	1.4182*	1.3040
IMCL (ours)	1.6159	1.4708	1.3760	1.2906

** means IMCL outperformed the corresponding approach significantly in paired t-test with Holm's p-value adjustment ($\alpha = 0.05$).*

4.3.1 Effect of Weight α

To analyze the sensitivity of IMCL to the weight α in the calculation of the pseudo-labels, we fixed each τ_t to $e_t/2$ and conducted experiments for $\alpha = \{0.1, 0.5, 0.9\}$. The average results on all three emotion dimensions of each dataset are shown in Fig. 4. IMCL almost always outperformed IMAL on RMSE, regardless of α, though their CCs were similar. Again, this may be due to the fact that only the RMSE was explicitly considered in the training of the regression models $\{f_t\}_{t\in\mathcal{T}}$.

For different α, the performance of IMCL was quite stable, especially for the RMSE. A closer look may reveal that generally a smaller α resulted in slightly better performance, especially when K was small. This is because when the size of the labeled sample set is small, kNN can better fit the sparse conditional label distribution. As the number of labeled sample increases, the label distribution becomes more complex and overwhelms the fitting capability of kNN. Hence, the label estimator g_t is more likely to help improve the performance of f_t, and the estimated conditional label should be assigned a larger weight, corresponding to a smaller α. These observations suggest that maybe an adaptive α should be used for better performance.

4.3.2 Effect of Threshold $\boldsymbol{\tau}$

We also conducted experiments with $\boldsymbol{\tau} = \{\boldsymbol{e}, \boldsymbol{e}/2, \boldsymbol{e}/5\}$ [where each $e_t \in \boldsymbol{e}$ was the training RMSE of the RR model in Task t] while fixing $\alpha = 0.5$. The average results on all three emotion dimensions are shown in Fig. 5. Again, IMCL almost always outperformed IMAL on RMSE, regardless of α, though their CCs were similar.

For different $\boldsymbol{\tau}$, the performance of IMCL on VAM and IAPS was quite stable. A closer look may reveal that generally a larger $\boldsymbol{\tau}$ resulted in slightly better performance, especially when K was small. This is because when K was small (the number of labeled samples was small), it is more beneficial to make use of more pseudo-labeled samples (corresponding to a larger $\boldsymbol{\tau}$), though they may be noisy. However, as K increased, g_t was not able to fit the conditional label distribution well. As a result, the benefit of the pseudo-labeled samples gradually vanished, and using too many such samples may even hurt the performance, as more clearly demonstrated on IEMOCAP. These observations suggest that maybe an adaptive $\boldsymbol{\tau}$ could be used for better performance.

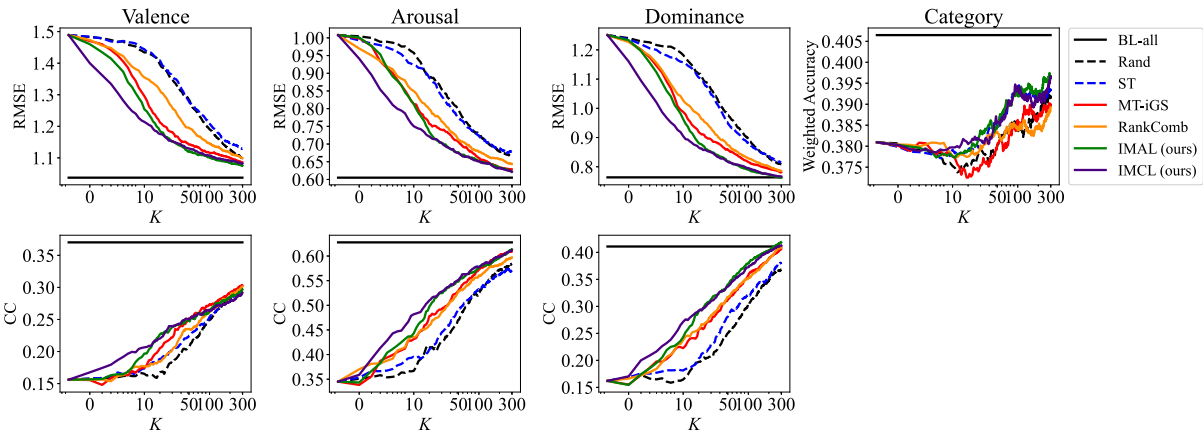

Fig. 3. Average performance of different sample selection algorithms on IEMOCAP in SECE (valence-arousal-dominance estimation and also emotion classification). Generally our proposed IMCL achieved the best performance, and IMAL the second best. K is the number of samples selected by different strategies to be manually annotated, in addition to the initial $d+1$ randomly selected labeled samples.

4.4 Discussion

The proposed IMCL may be explained as utilizing the prediction inconsistency from different models, f_t and g_t, to measure the informativeness of unlabeled samples and selecting the most informative ones for annotation. From this aspect, it resembles the model disagreement-based AL approaches. Typically, the classical QBC [18] constructs a committee of several base learners that trained on different subsets of labeled samples and selects the samples with maximum disagreement among the base learners. In multi-view learning, Muslea et al. [29] proposed to annotate samples that have inconsistent predictions from models in multiple views with different disagreement measure strategy.

However, IMCL has different motivation from these approaches: g_t represents the conditional label distribution of X^L, whereas f_t is the task learner; the inconsistency between f_t and g_t measures the label diversity a sample could bring to X^L. Whereas in conventional committee-based approaches, every model is a task learner, and the disagreement among them is usually supposed to be a measure of the prediction uncertainty.

5 Conclusion and Future Research

Discrete categories and continuous dimensional primitives are common emotion representations. To recognize categorical and dimensional emotions, usually a large number of labeled samples are required to train the classification and regression models. Manually labeling the samples in multi-task emotion recognition is costly and time-consuming, due to the ambiguity and subjectiveness of emotions. Multi-task AL can be used to save the labeling effort in MDEE and SECE and it can be further integrated with SSL to improve the performance of emotion recognition models with limited labeled samples.

TABLE 5
Average RMSEs of the Three Emotion Dimensions at the 5th, 10th, 20th and 50th Iterations, on IEMOCAP in SECE

Iteration	5	10	20	50
Rand	1.1441*	1.1087*	1.0350*	0.9210*
ST	1.1430*	1.1043*	1.0403*	0.9247*
MT-iGS	1.0742*	0.9845*	0.8993*	0.8377*
RankComb	1.0852*	1.0265*	0.9481*	0.8595*
IMAL (ours)	1.0616*	0.9696*	0.8796*	0.8163
IMCL (ours)	0.9973	0.9325	0.8691	0.8174

* *means IMCL outperformed the corresponding approach significantly in paired t-test with Holm's p-value adjustment* ($\alpha = 0.05$).

This paper has proposed an inconsistency measure that can indicate the difference between the labels estimated from the feature space and label distribution of the labeled dataset. It is then used in multi-task AL to select the most inconsistent sample to label, and in self-training SSL to select the least inconsistent samples to assign pseudo-labels. Experiments on three popular affective computing datasets demonstrated the effectiveness of our proposed IMCL, which integrates AL and SSL. It generally outperformed a state-of-the-art single-task AL approach and two multi-task AL approaches.

Our research can be improved or extended in the following ways:

1) Our proposed inconsistency measure utilizes the relationship among multiple tasks and is used as an informativeness indicator in AL and SSL. There are other important factors that could be considered in AL for regression [9], e.g., feature diversity and representativeness. They are complementary to label diversity and hence can be considered simultaneously for better performance. Our future research will integrate inconsistency with feature diversity and representativeness.
2) In our current approach, the first $d+1$ labeled samples are randomly selected. This initialization can also be optimized. For example, in [9] it has been shown that the representativeness and feature diversity criteria can be used for better initialization in AL for regression. Similar strategies will also be considered in our future research.
3) In our experiments, we fixed the hyper-parameters α and set the sample selection rules in SSL heuristically. Though IMAL and IMCL are not very sensitive to them, they do have some impact on the performance. Our future research will consider adaptive α and sample selection threshold/percentage in SSL with K, as suggested in Section 4.3. Additionally, the

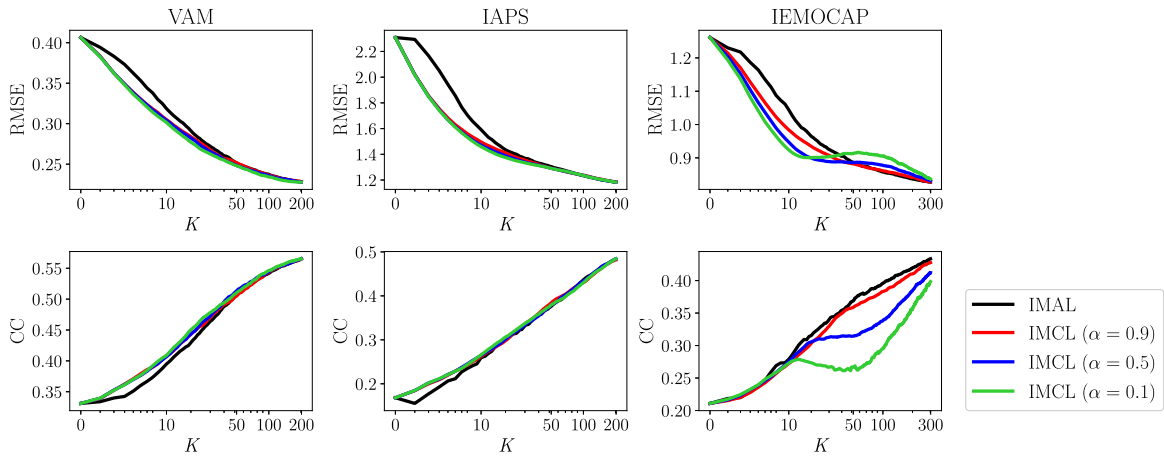

Fig. 4. Average RMSEs and CCs of IMCL on the three emotion dimensions in MDEE using different α. K is the number of samples selected by different strategies to be manually annotated, in addition to the initial $d + 1$ randomly selected labeled samples.

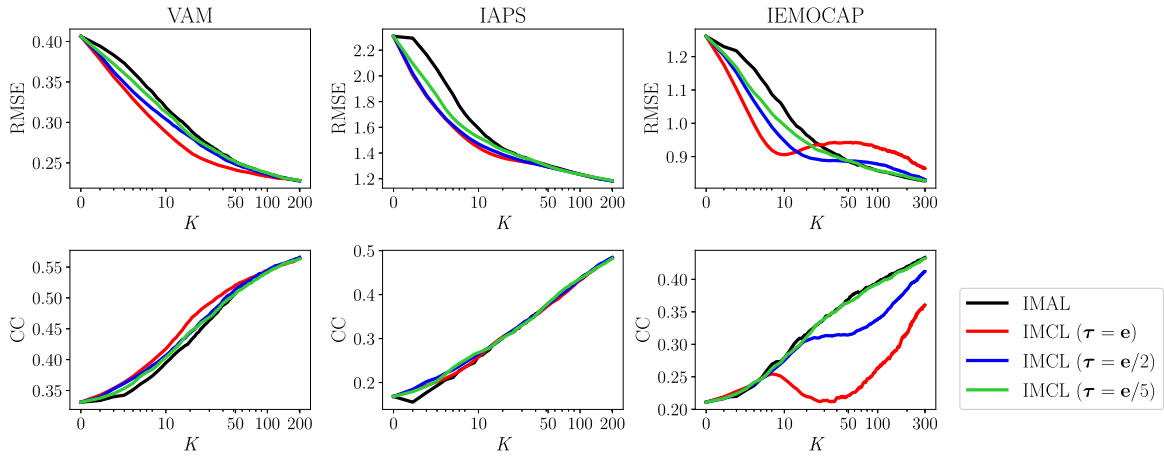

Fig. 5. Average RMSEs and CCs of IMCL on the three emotion dimensions in MDEE using different τ. K is the number of samples selected by different strategies to be manually annotated, in addition to the initial $d + 1$ randomly selected labeled samples.

models used in f_t and g_t may also change with K to better fit the distribution of dataset.

4) This paper only considered the inconsistency in MDEE and SECE. There also exist other multi-task scenarios in affective computing, e.g., determining both emotions and paralinguistics [30] in speech simultaneously. How to extend the inconsistency measure to such scenarios is another interesting problem.

REFERENCES

[1] J. Tao and T. Tan, "Affective computing: A review," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2005, pp. 981–995.

[2] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[3] R. E. Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2004, Art. no. 154.

[4] D. Wu, T. D. Parsons, E. Mower, and S. Narayanan, "Speech emotion estimation in 3D space," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2010, pp. 737–742.

[5] D. Wu et al., "Optimal arousal identification and classification for affective computing: Virtual Reality Stroop Task," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 109–118, Jul.–Dec. 2010.

[6] P. Ekman et al., "Universals and cultural differences in the judgments of facial expressions of emotion," *J. Pers. Soc. Psychol.*, vol. 53, no. 4, 1987, Art. no. 712.

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

[8] A. Mehrabian, *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Cambridge, MA, USA: Oelgeschlager, Gunn & Hain, 1980.

[9] D. Wu, "Pool-based sequential active learning for regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1348–1359, May 2019.

[10] D. Wu and J. Huang, "Affect estimation in 3D space using multi-task active learning for regression," *IEEE Trans. Affect. Comput.*, vol. 13, no. 41, pp. 16–27, Jan–March. 2022.

[11] D. Wu, "Active semi-supervised transfer learning (ASTL) for offline BCI calibration," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2017, pp. 246–251.

[12] G. Muhammad and M. F. Alhamid, "User emotion recognition from a larger pool of social network data using active learning," *Multimedia Tools Appl.*, vol. 76, no. 8, pp. 10 881–10 892, 2017.

[13] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 275–278.

[14] W. Han, H. Li, H. Ruan, L. Ma, J. Sun, and B. W. Schuller, "Active learning for dimensional speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 2841–2845.

[15] M. Abdelwahab and C. Busso, "Active learning for speech emotion recognition using deep neural network," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2019, pp. 1–7.

[16] D. Wu, C.-T. Lin, and J. Huang, "Active learning for regression using greedy sampling," *Inf. Sci.*, vol. 474, pp. 90–105, 2019.

[17] W. Cai, Y. Zhang, and J. Zhou, "Maximizing expected model change for active learning in regression," in *Proc. Int. Conf. Data Mining*, 2013, pp. 51–60.
[18] T. RayChaudhuri and L. G. Hamey, "Minimisation of data collection by active learning," in *Proc. IEEE Int. Conf. Neural Netw.*, 1995, pp. 1338–1341.
[19] X. Jiang, L. Meng, and D. Wu, "Multi-task active learning for simultaneous emotion classification and regression," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2021, pp. 1947–1952.
[20] R. Reichart, K. Tomanek, U. Hahn, and A. Rappoport, "Multi-task active learning for linguistic annotations," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2008, pp. 861–869.
[21] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 115–126, Jan. 2015.
[22] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2008, pp. 865–868.
[23] D. Wu, T. D. Parsons, and S. S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010.
[24] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Technical manual and affective ratings," *NIMH Center Study Emotion Attention*, vol. 1, pp. 39–58, 1997.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Image Net: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[27] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
[28] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.
[29] I. Muslea, S. Minton, and C. A. Knoblock, "Active learning with multiple views," *J. Artif. Intell. Res.*, vol. 27, pp. 203–233, 2006.
[30] B. Schuller et al., "Paralinguistics in speech and language – State-of-the-art and the challenge," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 4–39, 2013.

Yifan Xu received the BE degree in automation from the Huazhong University of Science and Technology, Wuhan, China, in 2018, where she is currently working toward the PhD degree in artificial intelligence with the School of Artificial Intelligence and Automation. Her research interests include affective computing, brain–computer interfaces, and machine learning.

Yuqi Cui received the BE degree in electronic information engineering, and the PhD degree in control science and engineering both from the Huazhong University of Science and Technology, Wuhan, China, in 2017 and 2022, respectively. His research interests include machine learning, fuzzy systems, and brain-computer interfaces.

Xue Jiang received the BE degree in communication engineering from Southwest University, Chongqing, China, in 2019. She is currently working toward the PhD degree in control science and engineering with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. Her research interests include brain-computer interfaces and machine learning.

Yingjie Yin received the PhD degree in pattern recognition and artificial intelligence from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2016. He was an assistant professor with the Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, from 2016 to 2018, and an associate professor from 2018 to 2020. Between 2017 and 2019, he was also a Hong Kong Scholar with the Department of Computing, Hong Kong Polytechnic University, Hong Kong, China. He joined Ant Group, in 2020 and is now a staff engineer of the Biometrics Technology Research Group. His research interests include computer vision and pattern recognition.

Jingting Ding received the PhD degree from Zhejiang university, China, in 2012. He joined Ant Group, in 2015, and is now a staff engineer. His main research interest is deep learning for computer vision, focusing particularly on biometrics and security, including face liveness, adversarial examples and Deepfake.

Liang Li received the PhD degree in pattern recognition and artificial intelligence from the Institute of Automation, Chinese Academy of Sciences, in 2007. He is a senior staff engineer with IOT business unit of Ant Group. Between 2007 and 2014, he was a senior researcher with Sony China Research Lab, and led multiple projects on image recognition for Sony's digital camera/TV/Playstation product line. He joined Ant Group, in 2014 and is now director of the biometrics technology research group, in charge of identity authentication for various financial applications. His research interests include biometrics, computer vision, and pattern recognition.

Dongrui Wu (Senior Member, IEEE) received the BE degree in automatic control from the University of Science and Technology of China, in 2003, the ME degree in electrical engineering from the National University of Singapore, in 2005, and the PhD degree in electrical engineering from the University of Southern California, in 2009. He is now a professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China, and deputy director of the Key Laboratory of Image Processing and Intelligent Control, Ministry of Education. His research interests include affective computing, brain-computer interface, computational intelligence, and machine learning. He has more than 190 publications, including more than 60 IEEE Transactions papers. He received the IEEE Computational Intelligence Society Outstanding PhD Dissertation Award, in 2012, the IEEE Transactions on Fuzzy Systems Outstanding Paper Award, in 2014, the NAFIPS Early Career Award, in 2014, the IEEE Systems, Man and Cybernetics (SMC) Society Early Career Award, in 2017, the USERN Prize in Formal Sciences, in 2020, the IEEE Transactions on Neural Systems and Rehabilitation Engineering Best Paper Award, in 2021, and the Chinese Association of Automation Early Career Award, in 2021. His team won the First Prize in China BCI Competition in three successive years (2019-2021). He is a BoG member and associate vice president for Human-Machine Systems of the IEEE SMC Society, and editor-in-chief of its *eNewsLetter*. He will be the editor-in-chief of the *IEEE Transactions on Fuzzy Systems*, in 2023.