# Improving Representation with Hierarchical Contrastive Learning for Emotion-cause Pair Extraction

Guimin Hu, Yi Zhao, Guangming Lu

**Abstract**—Emotion-cause pair extraction (ECPE) aims to extract emotions and their corresponding cause from a document. The previous works have made great progress. However, there exist two major issues in existing works. First, most existing works mainly focus on the semantic relation between the emotion clause and cause clause, ignoring their inner statistical relation in representation space. Second, the existing works are sensitive to the relative position between the emotion clause and cause clause, which damages the model's robustness. To address the two issues, we propose a hierarchical contrastive learning framework (HCL-ECPE), which hierarchically performs contrastive learning on representation from two levels. The first level is inter-clause contrastive learning (ICCL), which performs between emotion clause and cause clause through mutual information maximization. The second level is intra-pair contrastive learning (IPCL), which performs between clause representation and pair representation through contrastive predictive coding (CPC). HCL-ECPE integrates ICCL and IPCL modules to explore the statistical relations between the emotion clause, cause clause, and their constructed emotion-cause pair from the perspective of mutual information, thereby improving the model performance and robustness. Experimental results on two public datasets, ECPED and RECCON, demonstrate that HCL-ECPE outperforms the most competitive baselines. Furthermore, ICCL and IPCL are orthogonal to the existing model, and introducing them into the current models updates state-of-the-art performance.

**Index Terms**—Emotion-cause pair extraction, hierarchical contrastive learning, mutual information, contrastive predictive coding.

## 1 Introduction

Sentiment analysis and emotion recognition are popular topics in the field of understanding human behaviors and analyzing human intentions [1], [2]. Recently, some works on emotion cause analysis have risen, which focus on the cause behind a specific emotion. Emotion cause analysis [3], [4], [5], [6] is critical research and attracts recently lots of attention, gradually evolving into the emotion cause extraction (ECE) task and emotion-cause pair extraction (ECPE) task. ECE aims to determine which clauses contain the causes for the given emotion [7], [8], [9], [10]. However, ECE task ignores the fact that there exists mutually indicative impacts between emotion and cause and it also needs annotated emotions before cause clause extraction [11]. The ECPE task addresses this problem. Different from ECE, ECPE [11] aims to extract the emotions and their corresponding causes without requiring the emotion annotation in advance. As shown in the following example, the emotion clause $c_2$ and its cause clause $c_1$ are extracted from a document and form an emotion-cause pair, i.e., $(c_2, c_1)$:

**Example.** *Ever since his parents divorced ($c_1$), he has been depressed ($c_2$), and even wants to drop out of school ($c_3$).*

Early works of ECPE mainly focus on joint learning between emotion clause extraction, cause clause extraction, and emotion-cause pair extraction. For example, [11] proposes a two-step method, in which the first step is to extract emotion and cause clauses simultaneously. The second step is to pair the extracted clauses and filter out the non emotion-cause pairs from them. Nevertheless, this two-step setting may lead to error propagation from the first to the second step due to its pipeline architecture. To address this issue, recent works [3], [5], [12], [13], [14], [15], [16] integrate the emotion clause extraction, cause clause extraction and emotion-cause pair extraction into a unified architecture, and jointly train the tasks through multi-task learning. [4], [17] assume that each clause in the document can be viewed as an emotion clause or cause clause, then correspondingly extract their cause clauses and emotion clauses based on this assumption. Some pieces of works [3], [12], [18] convert ECPE as a sequence labeling problem to extract emotion-cause pairs, such as [12] takes the relative distance between the emotion clause and cause clause as part of labeling tag, [3] incorporates the emotional type into the labeling tag to deal with the cause clauses under different emotional type, and [18] utilizes a content part and a pairing part to identify the emotion/cause clauses and emotion-cause pair, respectively. Moreover, some advancing ECPE works use competitive backbones to incorporate more information, such as the knowledge base [5], graph neural networks [13], [14], [19], and Transformer-based models [20] to solve ECPE task. Some works explore the inter-clause or intra-clause knowledge. For

*This work was supported in part by the NSFC fund (NO. 62176077), in part by the Shenzhen Key Technical Project (JSGG20220831092805009), in part by the Guangdong International Science and Technology Cooperation Project (NO. 2023A0505050108), in part by the Shenzhen Fundamental Research Fund (NO. JCYJ20210324132210025), in part by the University Innovative Team Project of Guangdong, China under Grant No. 2022KCXTD039. (Corresponding author:Yi Zhao).*

*Guimin Hu and Guangming Lu are with the School of Computer Science, and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China, (e-mail: rice.hu.x@gmail.com; luguangm@hit.edu.cn).*
*Yi Zhao is with the School of Science, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China, (e-mail: zhao.yi@hit.edu.cn)*

example, [21] captures the fine-grained semantic cues of each clause, and [22] exploits external sentiment knowledge, intra-clause syntactic dependency, and inter-clause consistency to model local and global semantics.

Although the existing works of ECPE have achieved significant progress, there are two issues needed to be addressed among these works. One is the existing works focus on the relations between emotion clause and cause clause by using deep networks like attention mechanism [9], Transformer [23] and commonsense knowledge [10] to improve model performance, ignoring the inner statistical relation between them. In the work of [24], they found the inner statistical relation between the emotion clause and its cause clause, which denotes the dependencies between the emotion clause and cause clause in their representation space. The other issue is the sensitivity to the changing of relative position between the emotion clause and cause clause, which may damage the model's robustness. The work [24] had probed the feasibility of extracting emotion-cause pairs based on the mutual information between their emotion clause and cause clause. Meanwhile, the work [24] also demonstrates that the mutual information of the emotion-cause pair is less vulnerable to the changes of the relative distance [25]. Inspired by this work, we propose a hierarchical contrastive learning framework (HCL-ECPE) to solve the ECPE task. Hierarchical contrastive learning is widely applied to the multimodal sentiment analysis community [26], and these works aim to enhance the interaction between modalities by introducing contrastive learning loss as the model's regularizer.

In this work, HCL-ECPE implements inter-clause contrastive learning (ICCL) and intra-pair contrastive learning (IPCL) by incorporating the recently proposed deep infomax (DIM) [27] and noise-contrastive estimation framework [28] into the model, respectively. Different from the work of multimodal sentiment analysis, HCL-ECPE focuses on the use of mutual information to alleviate the dependency of the model on relative position between emotion clause and cause clause, and mines the statistical relation between them. First, introducing mutual information can alleviate the information loss caused by the long-distance dependency between the emotion clause and the cause clause. Second, the extraction framework based on mutual information can improve the model's robustness. This setting allows the model to focus on the deep causal relation, not just the semantic relation between the emotion clause and cause clause, thereby robustly extracting emotion-cause pairs from a document.

HCL-ECPE hierarchically regularizes representation learning with contrastive loss from inter-clause and intra-pair perspectives. Inter-clause relation denotes the relation between the emotion and cause clauses in the same emotion-cause pair, aiming to boost the clause representation learning by considering their relation in embedding space. In contrast with inter-clause relation, intra-pair relation represents the relation between emotion clause/cause clause and an emotion-cause pair composed of them, which aims to boost the emotion-cause pair representation learning by adjusting the pairing path from emotion clause to emotion-cause pair or cause clause to emotion-cause pair.

The contributions are summarized:

- We propose a hierarchical contrastive learning framework (HCL-ECPE) for the ECPE task. HCL-ECPE hierarchically performs inter-clause and intra-pair contrastive learning for sufficient representation learning.
- Up to our knowledge, it is the first work to leverage contrastive learning to model the hierarchical relations among the emotion clause, cause clause, and emotion-cause pair. Besides, the proposed inter-clause contrastive learning (ICCL) and intra-pair contrastive learning (IPCL) modules are orthogonal to the existing work. Introducing ICCL and IPCL as regularizer terms can bring improvements for the existing models.
- Experiment results on two public datasets, ECPED and RECCON, show that HCL-ECPE outperforms most competitive baselines and further demonstrates the effectiveness of ICCL and IPCL modules.

## 2 Related Work

### 2.1 ECPE

ECE is firstly proposed in [7], in which a new corpus is constructed based on the Academia Sinica Balanced Chinese Corpus. Based on this corpus, [29] proposed a multi-label method and released two groups of linguistic features. [30] proposed to identify linguistic contexts to solve ECE automatically. Previous studies adapted words as the annotation granularity of emotion cause, resulting in incomplete semantics of extracted emotion cause. To address this issue, [8] released a clause-level Chinese emotion cause corpus and adopted an event-driven multi-kernel SVM model to extract the emotion cause clause. The Chinese emotion cause corpus is collected from SINA city news [1]. It is gradually used in many works to evaluate the performance of ECE as the benchmark. In recent years, [9], [10], [11], [31] adopted deep neural networks, like multiple-slot memory network [32], joint learning [33], co-attention network [9], Transformer network [23], and knowledge-based regularization [31], to solve ECE task. However, the ECE task faced with two shortcomings: 1) it requires emotion annotation before extracting the cause clause, and 2) it ignores the mutual indications between the cause clause and the triggered emotion clause.

Emotion-cause pair extraction (ECPE) [11] is the extension of emotion-cause extraction (ECE), which aims to extract the emotion clauses and their cause clause simultaneously from an unannotated document. ECPE addresses the two issues of ECE and its goal is to simultaneously extract the emotion clause and cause clause from an unannotated document and form emotion-cause pairs between the emotion clause and cause clause. Initially, [11] proposed a two-step method to extract and pair the emotion clauses and cause clauses, which may propagate the errors from the first to the second step due to the pipeline architecture structure. More recently, some pieces of work extracted emotion-cause pairs with unified frameworks to avoid this issue. For example, [14] scored candidate pairs and extracted the pairs with the top score as the emotion-cause pairs. [20] extended original

1. http://news.sina.com.cn/society/

Transformer [34] to 2D-Transformer to model the interactions between clause pairs. [4] assumed that each clause in the document is viewed as the emotion clause or cause clause separately to extract the corresponding cause clauses or emotion clauses based on this premise. Some pieces of literature [3], [12], [18], [35], [36] viewed the ECPE task as a sequence labeling problem and extracted emotion-cause pairs by labeling emotion clauses, cause clauses and their pairing relation simultaneously. For example, [12] took the relative distance between the emotion clause and cause clause as a part of the label. [3] utilized the emotional type of emotion clause into the label to distinguish the cause clauses under different emotional types. [18] designed the content and pairing parts to redefine ECPE as a unified sequence labeling problem. [35] took the predicted distribution of auxiliary tasks to adjust the paired tagging distribution of the distances between the emotion clause and cause clause into a novel tagging scheme. Besides, [36] proposed a dual-questioning attention network, which viewed the candidate's emotions and causes as two individual questions to extract their causes or emotions from the context independently. Although existing works have made significant progress, some issues need to be addressed. We summarize the issues into two aspects: 1)the existing works ignore the inner statistical relation between the emotion clause and the cause clause, and 2)the performances of existing models are easily affected by the relative position between the emotion clause and cause clause. To address the two issues, we propose a hierarchical contrastive learning framework (HCL-ECPE), aiming to mine the internal relationships among the emotion clause, cause clause, and their composed emotion-cause pair in the representation space.

### 2.2 Contrastive Learning

Recently, contrastive learning [37] has made significant advances in many deep learning areas, especially in representation learning. Its principle is clear, following the idea that an anchor and its positive sample should be pulled close. In contrast, the anchor and negative samples should be pushed apart in feature space by viewing samples from multiple views. In practice, contrastive learning methods benefit from the comparison between positive and negative sample representation learning [38], [39]. Mutual information (MI) is a concept from information theory that estimates the relationship between variable pairs and can also be viewed as a member of contrastive learning. Mutual information is a reparameterization invariant measure of the amount of information obtained from one random variable through another, measuring the statistical dependence between variables. Mutual information estimation can be grouped into explicit estimation and approximate estimation, where explicit estimation denotes the non-parametric methods [40]. Approximate estimation denotes the parametric methods [41]. In the NLP field, [42] reformulated the original question in another expression through the mutual information maximization between question and answer. [43] added the mutual information (MI) term between the input and its latent variable to the objective of the variational auto-encoder (VAE) to improve representation learning. [44] leveraged the upper bound of mutual information to induce style and content embedding into two independent low-dimensional spaces, thereby maintaining consistency. [26] maximized the mutual information in uni-modal input pairs (inter-modality) and between multi-modal fusion results to incorporate task-related information into multi-modal representation. [45] proposed a regularizer in reading comprehension systems by maximizing mutual information among a passage, a question, and its answer to learn their correlates. [46] used the KL differences and mutual information as the constraints to ensure consistency between the structure and semantics. Mutual information describes the high-order dependence among variables, and mutual information of two random variables $X$ and $Y$ is defined as:

$$\mathrm{MI}(X,Y) = D_{KL}(p(X,Y)||p(X)p(Y)) \tag{1}$$

where $D_{KL}$ is the Kullback-Leibler (KL) divergence between the joint distribution $p(X,Y)$ and the product of marginals of $X$ and $Y$. Since the exact computation of MI is intractable, we use a neural approximation method MINE. MINE (mutual information neural estimation) [47] estimates the mutual information by training a classifier to distinguish whether samples come from the joint distribution of $X$ and $Y$ or from the product of their marginal distributions. MINE uses a lower-bound to the MI based on the Donsker-Varadhan (DV) [48] representation of the KL-divergence as the approximate estimation of mutual information:

$$\mathrm{MI}(X,Y) = \mathbb{E}_{\mathbb{P}}[g(x,y))] - \log(\mathbb{E}_{\mathbb{N}}[e^{g(x,\tilde{y})}]) \tag{2}$$

where $\mathbb{E}_{\mathbb{P}}$ and $\mathbb{E}_{\mathbb{N}}$ denote the expectation over positive and negative samples, respectively, and $g(x,y)$ is a discriminator function. [27] mentioned that the DV representation shown in the MINE is the strong bound of mutual information while maximizing MI does not require the precise value. Based on this finding, [27] proposed to use Jensen-Shannon divergence (JS) to calculate MI, which can be efficiently implemented by the cross-entropy (BCE) loss:

$$\mathrm{MI}(X,Y) = \mathbb{E}_{\mathbb{P}}[g(x,y))] - \mathbb{E}_{\mathbb{N}}[\log(1-g(x,\tilde{y})] \tag{3}$$

Contrastive predictive coding (CPC) [49] is also a member of the contrastive learning family. CPC learns such representations by using a powerful autoregressive model to predict the future in the potential space, which induces the potential space to capture the most helpful information for predicting future samples. CPC combines autoregressive modeling and noise comparison estimation, aiming to learn abstract representation in an unsupervised way. When predicting future information, CPC encodes the target future and context into a compact distributed vector representation (through nonlinear learning mapping), which maximizes the mutual information between the original signal $x$ and its context $c$, which is defined as:

$$\mathrm{MI}(x,c) = \sum_{x,c} p(x,c)\log\frac{p(x|c)}{p(x)} \tag{4}$$

We extract the underlying latent variables by maximizing the mutual information between the encoded representations (bounded by the MI between the input signals). In this work, we adapt contrastive predictive coding to predict representations between emotion (or cause) clause and emotion-cause clause pairs so that more informative features

of clause presentation can be passed to the representation of clause pairs.

## 3 METHODOLOGY

### 3.1 Motivation

The preliminary work [24] found two phenomena for the emotion-cause pair extraction task: 1) the mutual information of the emotion-cause pair is significantly greater than the mutual information of non-emotion-cause pair, and 2) the mutual information of emotion-cause pair is less affected by the relative distance of its emotion clause and cause clause. Inspired by the two findings, we present a hierarchical contrastive learning framework (HCL-ECPE) to regularize clause and pair representation learning. On one hand, the model with mutual information maximization can capture the profound statistical correlation between the emotion clause and cause clause. On the other hand, the model with mutual information maximization alleviates the model's vulnerability to relative position changes between the emotion clause and cause clause [25]. The previously published works usually focus on the semantic information of intra-clause or inter-clause. Different from these works, HCL-ECPE uses contrastive loss as the regularization term to constrict the representations of emotion clauses and cause clauses in feature space. HCL-ECPE emphasizes the inter-clause and intra-pair statistical relation in terms of mutual information, rather than relative position that the previous works widely used.

### 3.2 Task Definition

Consider that a document $d = \{c_1, c_2, ..., c_{|d|}\}$ contains $|d|$ clauses, and each clause $c_i = \{w_{i1}, w_{i2}, ..., w_{in}\}$ contains $n$ words. The target of ECPE is to extract the emotion-cause pairs in the form of $\{\cdots, (c_{emo}, c_{cau})^j, \cdots\}$ from an unannotated document $d$, where $c_{emo}$ and $c_{cau}$ are emotion clause and cause clause for $j$th emotion-cause pair, respectively. We use a pair set $R_d$ to represent the extracted emotion-cause pairs from document $d$:

$$R_d = \{(c_{emo^1}, c_{cau^1}), ..., (c_{emo^l}, c_{cau^l})\} \tag{5}$$

where superscript $l$ denotes that the document $d$ has $l$ emotion-cause pairs.

### 3.3 Overall Framework

This section describes the architecture of HCL-ECPE, as shown in Figure 1. HCL-ECPE contains the hierarchical encoder, prediction layer, and two contrastive learning modules: inter-clause contrastive learning (ICCL) and intra-pair contrastive learning (IPCL). We first obtain emotion-specific and cause-specific clause representation by two individual hierarchical encoders. After obtaining clause representation, the ICCL module tunes the clause representation by maximizing the mutual information between the emotion clause and cause clause, aiming to capture the causal bond between them. Additionally, we obtain initial clause pair representations by packaging clause representations and then perform IPCL to learn an encoder that generates a path from the emotion or cause clause representation to the emotion-cause pair representation. IPCL optimizes clause representation with the opposite generation path, from clause pair representation to clause representation. ICCL aims to learn the mutual impacts of emotions and causes, while IPCL seeks to build the learning path between the emotion-cause pair and its emotion clause or cause clause.

### 3.4 Hierarchical Encoder

A hierarchical encoder models the hierarchical structure of the document and it contains a word-level encoder and clause-level encoder. Word-level encoder aims to make each word of the clause sequence aware of its context and capture the semantic relation among words. First, each word is mapped into a vector after a word embedding layer, and then they are fed into Bi-LSTM (Bidirectional Long Short-Term Memory) [50] to model the sequence feature of the clause. We can obtain the hidden state of $w_{ij}$ by concatenating forward hidden state $\overrightarrow{h_{ij}}$ and backward hidden state $\overleftarrow{h_{ij}}$:

$$h_{ij} = [\overrightarrow{h_{ij}}, \overleftarrow{h_{ij}}], \tag{6}$$

where $[\cdot, \cdot]$ is the concatenation operation. An attention layer [51] is adopted to aggregate the hidden states of each word and produce a state vector $h_i$ for clause $c_i$.

$$\begin{aligned} u_{ij} &= \tanh(W_w h_{ij} + b_w)^{\mathsf{T}} u_u \\ \alpha_{ij} &= \frac{\exp(u_{ij})}{\sum_j \exp(u_{ij})} \\ h_i &= \sum_{j=1}^{n} \alpha_{ij} h_{ij} \end{aligned} \tag{7}$$

where $W_w$ and $u_u$ are learnable weight matrix and vector, respectively. $\alpha_{ij}$ is the attention weight, showing the importance of word $w_{ij}$ to clause $c_i$. The clause representation $h_i$ is computed as a weighted sum of the word hidden states in the clause based on the weight $\alpha_{ij}$. The clause-level encoder models the sequential relation of a clause in a document by feeding them into Bi-LSTM and then generate the representations $\{r_1, r_2, ..., r_{|d|}\}$. Note that we also use BERT [52] (BERT-base Chinese [2]) as an alternative clause encoder to produce clause representations. We use two individual feed-forward networks (FFNs) to transform each clause representation to the emotion-specific representation $r_i^e$ and cause-specific representation $r_j^c$, respectively:

$$\begin{aligned} r_i^e &= \text{FFN}(r_i) \\ r_j^c &= \text{FFN}(r_j) \end{aligned} \tag{8}$$

### 3.5 Inter-clause Contrastive Learning (ICCL)

The core idea of contrastive learning [53] is to pull the positive sample pairs close and push the negative sample pairs far from each other in the feature space. We perform contrastive learning between clauses, aiming to maximize the mutual information between the emotion clause and cause clause representations, thereby enhancing their correlation. For each training batch, it consists of $M$ emotion-cause pairs represented by $\{(c^e, c^c)^1, ..., (c^e, c^c)^M\}$. Based on the clause representations specific to emotion and cause,

2. https://github.com/google-research/bert

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3391854



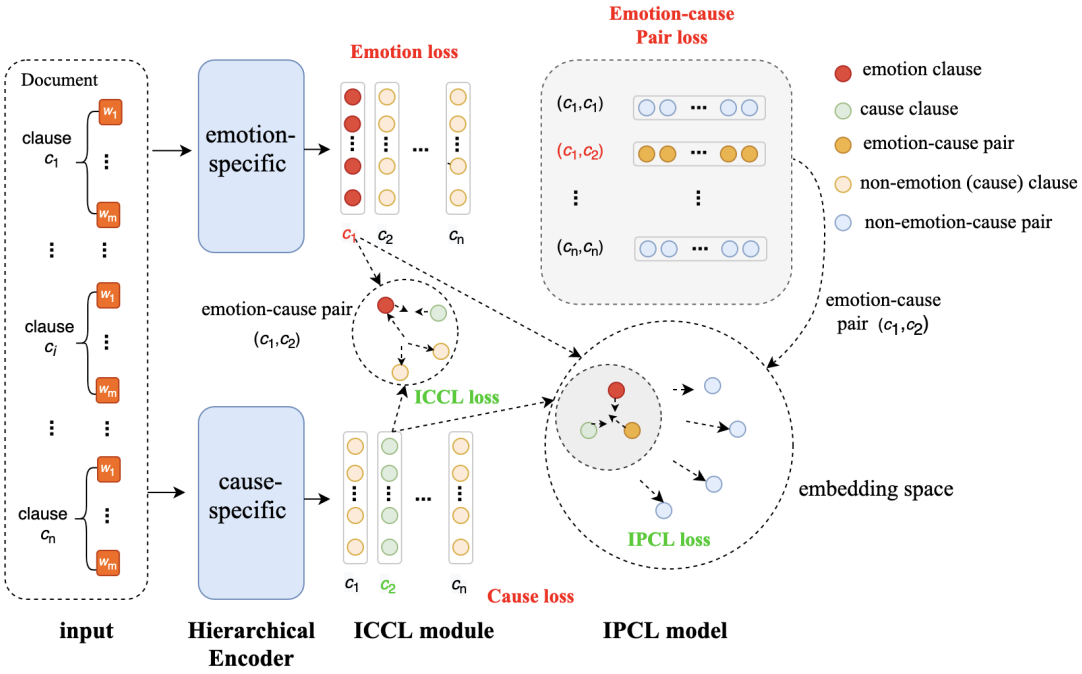



Fig. 1. The overview of HCL-ECPE. Inter-clause and intra-pair contrastive learning are performed on the emotion/cause-specific representation and emotion-cause pair representation, respectively. The objective loss contains a task loss and contrastive learning loss, which are marked with red and green dashed boxes, respectively.

we obtain a tuple $(r^e_{i_k}, r^c_{j_k})^M$ composed of the emotion clause representation and cause clause representation. The subscripts $i^e_k$ and $j^c_k$ indicate the indexes of the emotion clause and cause clause, respectively, for the $k$th emotion-cause pair. Then we obtain a tuple set for the training batch $\{(r^e_{i_1}, r^c_{j_1})^1, ..., (r^e_{i_M}, r^c_{j_M})^M\}$ as the positive examples, i.e., the examples sampled from the joint distribution of emotion clause and cause clause. We generate negative examples by shuffling the emotion-cause pairs in the batch such that an emotion clause is randomly associated with a cause clause in the batch. Each sample has two negative cases. One is the representation tuple composed of the emotion clause and non-cause clause, and the other is the representation tuple composed of the non-emotion clause and cause clause. The tuple sets for the training batch $\{(\tilde{r}^e_{i_1}, r^c_{j_1})^1, ..., (\tilde{r}^e_{i_M}, r^c_{j_M})^M\}$ and $\{(r^e_{i_1}, \tilde{r}^c_{j_1})^1, ..., (r^e_{i_M}, \tilde{r}^c_{j_M})^M\}$ are used as the negative examples. Each sample has one positive case and two negative cases. Following the usage of mutual information in QA community [45], the mutual information estimation of clause pair $(c^e_i, c^c_j)$ is give by the bound:

$$\begin{aligned}\mathrm{MI}(c^e_i, c^c_j) &\geq \mathbb{E}_{r^e_i, r^c_j \sim \mathbb{P}}[\log g(r^e_i, r^c_j)] \\ &+ \frac{1}{2}\mathbb{E}_{\tilde{r}^e_i, r^c_j \sim \mathbb{N}}[\log g(\tilde{r}^e_i, r^c_j)] \\ &+ \frac{1}{2}\mathbb{E}_{r^e_i, \tilde{r}^c_j \sim \mathbb{N}}[\log g(r^e_i, \tilde{r}^c_j)]\end{aligned} \tag{9}$$

where $\mathbb{E}_{\mathbb{P}}$ and $\mathbb{E}_{\mathbb{N}}$ denote the expectation over positive and negative examples, respectively. Intuitively, the function $g(\cdot)$ acts like a binary classifier that discriminates whether the clause pairs $(c_i, c_j)$ is the positive sample or not. Specially, the expression of $g(\cdot)$ is given as follows:

$$g(x, y) = \mathrm{sigmoid}(x^T W y) \tag{10}$$

where $x$ and $y$ denote the representations of the emotion clause and cause clause, respectively. $W$ is the learnable matrix. Note that we set dropout for emotion and cause clause representations, and then feed them into $g(x, y)$.

The objective loss of ICCL $\mathcal{L}_{iccl}$ is given by with the lower bound of MI, i.e., $\overline{\mathrm{MI}}$:

$$\mathcal{L}_{iccl} = -\frac{1}{M}\sum_{i,j}\overline{\mathrm{MI}}(c^e_i, c^c_j) \tag{11}$$

### 3.6 Intra-pair Contrastive Learning (IPCL)

Different from inter-clause contrastive learning, the target of intra-pair contrastive learning (IPCL) is to learn a path from an emotion clause or a cause clause to their composed emotion-cause pair, which improves representation learning by exploiting informative features between them. Meanwhile, IPCL can adjust the relative position between the emotion clause or cause clause and emotion-cause pair in embedding space. IPCL aims to generate a clause pair encoder function $F$ that produces emotion-cause pair representation $p_{ij} = F(r^e_i, r^c_j)$ based on representations of emotion clause and cause clause. $F$ is a parameterized network (e.g., multilayer perceptron). Specifically, $p_{ij} = \mathrm{Relu}(W_f[r^e_i, r^c_j] + b_f)$. Assumed that we already have a generation path $F$ from $r^e_i$ and $r^c_j$ to $p_{ij}$, and we expect to construct $r^e_i$ and $r^c_j$ from $p_{ij}$ with the opposite path. Take the emotion clause representation $r^e_i$ as an example. We use a contrastive predictive coding (CPC) [49] score function that acts on the normalized prediction and truth vector to measure the

correlation of emotion clause and emotion-cause pair:

$$\overline{M_\phi(p_{ij})} = \frac{M_\phi(p_{ij})}{||M_\phi(p_{ij})||_2}, \overline{r_i^e} = \frac{r_i^e}{||r_i^e||_2}$$
$$s(r_i^e, M_\phi(p_{ij})) = \exp(\overline{r_i^e}(\overline{M_\phi(p_{ij})})^T) \quad (12)$$

where $M_\phi$ is a neural network (e.g., multilayer perceptron) with parameters $\phi$, aiming to generate emotion clause representation $r_i^e$ from the pair representation $p_{ij}$, i.e., $M_\phi(x) = \text{Relu}(W_\phi x + b_\phi)$. $||\cdot||_2$ is the Euclidean norm. As [49] did, we incorporate this score function into the noise-contrastive estimation framework [28] by treating all other clause representations $\tilde{R}^e$ in the same batch as negative samples:

$$\mathcal{L}^e_{ipcl}(p_{ij}, r_i^e) = -\mathbb{E}[\log \frac{s(r_i^e, M_\phi(p_{ij}))}{\sum_{r_l^e \in \tilde{R}^e} s(r_l^e, M_\phi(p_{lj})) + \sum_{r_i^e \in R^e} s(r_i^e, M_\phi(p_{ij}))}] \quad (13)$$

where $R^e$ denotes positive samples from emotion clause perspective. We also use a contrastive predictive coding (CPC) score function to measure the correlation of cause clause and emotion-cause pair. For the path from $r_j^c$ to $p_{ij}$, its noise-contrastive estimation is given as:

$$\mathcal{L}^c_{ipcl}(p_{ij}, r_j^c) = -\mathbb{E}[\log \frac{s(r_j^c, M_\phi(p_{ij}))}{\sum_{r_l^c \in \tilde{R}^c} s(r_l^c, M_\phi(p_{il})) + \sum_{r_j^c \in R^c} s(r_j^c, M_\phi(p_{ij}))}] \quad (14)$$

where $R^c$ and $\tilde{R}^c$ denotes positive samples and negative samples, respectively from the cause clause perspective. CPC scores the MI between the emotion clause (or cause clause) and the emotion-cause pair to ensure that the important information of emotion or cause clause representations is received by the emotion-cause pair representation. We use the representation of emotion-cause pair $p_{ij}$ to reversely optimize emotion or cause clause representations, so that the information passed from $\{r_i^e, r_j^c\}$ to $p_{ij}$ can be determined by $p_{ij}$. The loss function for the IPCL module is given by:

$$\mathcal{L}_{ipcl} = \beta^e \mathcal{L}^e_{ipcl}(p_{ij}, r_i^e) + \beta^c \mathcal{L}^c_{ipcl}(p_{ij}, r_j^c) \quad (15)$$

where $\beta^e$ denotes the hyperparameter of the contrastive loss between the emotion clause and emotion-cause pair. Similarly, $\beta^c$ denotes the hyperparameter of the contrastive loss between the cause clause and the emotion-cause pair.

### 3.7 HCL-ECPE Model

Except for auxiliary loss $\mathcal{L}_{iccl}$ and $\mathcal{L}_{ipcl}$, the objective loss of HCL-ECPE model contains emotion clause prediction, cause clause prediction, and emotion-cause pair prediction. The pair representation of $(c_i, c_j)$ is obtained based on $p_{ij} = F(r_i^e, r_j^c)$. Based on $r_i^e$, $r_j^c$ and $p_{ij}$, we can obtain the emotion prediction $\hat{y}_i^e$, cause prediction $\hat{y}_j^c$ and emotion-cause pair prediction $\hat{y}_{ij}^p$, respectively:

$$\hat{y}_i^e = \text{Softmax}(\text{FFN}(r_i^e))$$
$$\hat{y}_j^c = \text{Softmax}(\text{FFN}(r_j^c)) \quad (16)$$
$$\hat{y}_{ij}^p = \text{Softmax}(\text{FFN}(p_{ij}))$$

TABLE 1
The details of dataset ECPED.

| Number of items | Number | Percentage(%) |
|---|---|---|
| Document with one emotion-cause pair | 1,746 | 89.77 |
| Document with two or more emotion-cause pair | 199 | 10.23 |
| Emotion-cause pairs | 2,154 | - |
| Emotion clause | 2,085 | - |
| Cause clause | 2,142 | - |
| Average document length | 14.75 | - |

TABLE 2
The details of dataset RECCON.

| Number of items | Number |
|---|---|
| Dialogues | 1,122 |
| Utterances | 11,769 |
| Utterances annotated with emotion cause | 6,355 |
| Utterances that cater to background cause | 465 |
| Utterances where cause solely lies in the same utterance | 1,601 |
| Utterances where cause solely lies in the contextual utterances | 1,141 |
| Utterances where cause lies both in same and context utterances | 3,613 |

After obtaining prediction, the cross-entropy losses of emotion clause extraction and cause clause extraction are formalized as:

$$\mathcal{L}_t = \frac{1}{N}\sum_k^N \sum_i^{n_k} -[y_i^t \log \hat{y}_i^t + (1 - y_i^t)(1 - \log \hat{y}_i^t)] \quad (17)$$

where the superscript $t \in \{e, c\}$, $e$ and $c$ represent emotion clause extraction and cause clause extraction, respectively. $N$ is the number of training document, $n_k$ denotes the number of clauses in document $d_k$, $y_i^t$ is the real distribution of clause $c_i$ for task $t$. Moreover, the cross-entropy loss of emotion-cause pair extraction is given by:

$$\mathcal{L}_p = \frac{1}{N}\sum_k^N \sum_{i,j}^{m_k} -[y_{ij}^p \log \hat{y}_{ij}^p + (1 - y_{ij}^p)(1 - \log \hat{y}_{ij}^p)] \quad (18)$$

where $m_k$ denotes the number of clause pairs in document $k$. $y_{ij}^p$ is real distribution of clause pair $(c_i, c_j)$ for emotion-cause pair extraction.

During the training phase, we use the task loss with auxiliary hierarchical contrastive learning loss to train HCL-ECPE. The overall loss function of HCL-ECPE is given by:

$$\mathcal{L} = \mathcal{L}_e + \mathcal{L}_c + \mathcal{L}_p + \alpha \mathcal{L}_{iccl} + \beta \mathcal{L}_{ipcl} \quad (19)$$

where $\mathcal{L}_e$, $\mathcal{L}_c$ and $\mathcal{L}_p$ are the cross-entropy losses of emotion clause extraction, cause clause extraction, and emotion-cause pair extraction, respectively. $\alpha$ and $\beta$ are the hyperparameters of $\mathcal{L}_{iccl}$ and $\mathcal{L}_{ipcl}$, respectively. Furthermore, ICCL and IPCL modules are orthogonal to the existing work, and we can introduce them as the regularizer terms.

## 4 Experiments

### 4.1 Set Up

We conduct experiments on datasets ECPED [11] and RECCON [54] to evaluate the model performance. ECPED extends the benchmark dataset of the ECE task [8]. ECPED contains 1,945 Chinese documents. Each document has at least one or more emotion-cause pairs. It can be observed that the documents containing one emotion-cause pair and two or more emotion-cause pairs account for 89.77% and 10.23% in the benchmark dataset, respectively. Moreover,

TABLE 3
Experimental results of methods using hierarchical LSTMs as encoder on emotion-cause pair extraction (Pair Extraction in short), emotion extraction and cause extraction. The results with underline denote the previous SOTA performance.

| | Pair Extraction | | | Emotion Extraction | | | Cause Extraction | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| Inter-EC | 0.6721 | 0.5705 | 0.6128 | 0.8364 | 0.8107 | 0.8230 | 0.7041 | 0.6083 | 0.6507 |
| E2EECPE | 0.6478 | 0.6105 | 0.6280 | 0.8595 | 0.7915 | 0.8238 | 0.7062 | 0.6030 | 0.6503 |
| PairGCN | 0.6999 | 0.5779 | 0.6321 | 0.8587 | 0.7208 | 0.7829 | 0.7283 | 0.5953 | 0.6541 |
| LML | 0.6990 | 0.5960 | 0.6440 | 0.8810 | 0.7810 | 0.8260 | - | - | - |
| ECPE-2D | 0.6960 | 0.6118 | 0.6496 | 0.8512 | 0.8220 | 0.8358 | 0.7272 | 0.6298 | 0.6738 |
| SLSN-U | 0.6836 | 0.6291 | 0.6545 | 0.8406 | 0.7980 | 0.8181 | 0.6992 | 0.6588 | 0.6778 |
| RANKCP | 0.6698 | 0.6546 | 0.6610 | 0.8703 | 0.8406 | 0.8548 | 0.6927 | 0.6743 | 0.6824 |
| ECPE-MLL | 0.7090 | 0.6441 | 0.6740 | 0.8582 | 0.8429 | 0.8500 | 0.7248 | 0.6702 | 0.6950 |
| IE-CNN-CRF | 0.7149 | 0.6279 | 0.6686 | 0.8614 | 0.7811 | 0.8188 | 0.7348 | 0.5841 | 0.6496 |
| UTOS | 0.6911 | 0.6193 | 0.6524 | 0.8610 | 0.7925 | 0.8250 | 0.7189 | 0.6496 | 0.6802 |
| MGSAG | 0.7243 | 0.6507 | 0.6846 | 0.8721 | 0.7911 | 0.8287 | 0.7510 | 0.6713 | 0.7080 |
| EPO-ECPE | 0.7900 | 0.6021 | 0.6824 | **0.9780** | 0.7848 | 0.8702 | 0.7961 | 0.6039 | 0.6848 |
| HCL-ECPE | 0.6569 | 0.6534 | 0.6537 | 0.8614 | 0.7811 | 0.8134 | 0.6848 | 0.6587 | 0.6676 |
| E2EECPE(HCL) | 0.6683 | 0.6517 | 0.6632 | 0.8530 | 0.7961 | 0.8232 | 0.6663 | 0.6679 | 0.6712 |
| RANKCP(HCL) | 0.6851 | **0.6777** | 0.6782 | 0.8781 | 0.8114 | 0.8419 | 0.7121 | **0.6897** | 0.6910 |
| ECPE-MLL(HCL) | 0.6944 | 0.6765 | 0.6856 | 0.8815 | 0.8517 | 0.8620 | 0.7362 | 0.6844 | 0.7061 |
| UTOS(HCL) | 0.6820 | 0.6754 | 0.6781 | 0.8804 | **0.8523** | 0.8467 | 0.7216 | 0.6809 | 0.6950 |
| EPO-ECPE(HCL) | **0.7919** | 0.6205 | **0.7056** | 0.9711 | 0.8073 | **0.8789** | **0.7968** | 0.6631 | **0.7255** |

TABLE 4
Experimental results of methods using BERT as encoder on emotion-cause pair extraction, emotion clause extraction and cause clause extraction.

| | Pair Extraction | | | Emotion Extraction | | | Cause Extraction | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| PairGCN | 0.7692 | 0.6791 | 0.7202 | 0.8857 | 0.7958 | 0.8375 | 0.7907 | 0.6928 | 0.7375 |
| LMB | 0.7110 | 0.6070 | 0.6550 | 0.8990 | 0.8000 | 0.8470 | - | - | - |
| ECPE-2D | 0.7292 | 0.6544 | 0.6889 | 0.8627 | 0.9221 | 0.8910 | 0.7336 | 0.6934 | 0.7123 |
| RANKCP | 0.7119 | 0.7630 | 0.7360 | 0.9123 | 0.8999 | 0.9057 | 0.7461 | 0.7788 | 0.7615 |
| ECPE-MLL | 0.7700 | 0.7235 | 0.7452 | 0.8608 | 0.9191 | 0.8886 | 0.7382 | 0.7912 | 0.7630 |
| Transition | 0.7374 | 0.6307 | 0.6799 | 0.8716 | 0.8244 | 0.8474 | 0.7562 | 0.6471 | 0.6974 |
| Tagging | 0.7243 | 0.6366 | 0.6776 | 0.8196 | 0.7329 | 0.7739 | 0.7490 | 0.6602 | 0.7018 |
| UTOS | 0.7389 | 0.7062 | 0.7203 | 0.8815 | 0.8321 | 0.8559 | 0.7671 | 0.7320 | 0.7471 |
| Refinement | 0.7377 | 0.6802 | 0.7078 | 0.8593 | 0.7993 | 0.8282 | 0.7614 | 0.7039 | 0.7315 |
| EPO-ECPE | 0.7621 | 0.7519 | 0.7564 | 0.9787 | 0.9232 | 0.9500 | 0.7711 | 0.7543 | 0.7620 |
| UECA-Prompt | 0.7182 | **0.7799** | 0.7470 | 0.8475 | 0.9195 | 0.8816 | 0.7624 | 0.7916 | 0.7755 |
| A2Net | 0.7503 | 0.7780 | 0.7634 | 0.9067 | 0.9098 | 0.9080 | 0.7762 | 0.7920 | 0.7835 |
| HCL-ECPE | 0.6971 | 0.7637 | 0.7308 | 0.9081 | 0.9113 | 0.9092 | 0.7355 | 0.7755 | 0.7540 |
| HCL-ECPE(GCN) | 0.7337 | 0.7737 | 0.7511 | 0.9184 | 0.9239 | 0.9173 | 0.7687 | 0.7883 | 0.7712 |
| HCL-ECPE(Transformer) | 0.7371 | 0.7782 | 0.7608 | 0.9185 | 0.9286 | 0.9193 | 0.7889 | 0.7825 | 0.7846 |
| RANKCP(HCL) | 0.7325 | 0.7714 | 0.7512 | 0.9125 | 0.9065 | 0.9099 | 0.7645 | 0.7861 | 0.7722 |
| ECPE-MLL(HCL) | 0.7611 | 0.7523 | 0.7566 | 0.8781 | 0.9241 | 0.9051 | 0.7619 | 0.7969 | 0.7782 |
| UTOS(HCL) | 0.7420 | 0.7251 | 0.7363 | 0.8904 | 0.8417 | 0.8657 | 0.7686 | 0.7209 | 0.7418 |
| A2Net(HCL) | 0.7685 | 0.7771 | 0.7692 | 0.9048 | 0.9022 | 0.9011 | 0.7762 | **0.7974** | **0.7869** |
| EPO-ECPE(HCL) | **0.7914** | 0.7665 | **0.7756** | **0.9804** | **0.9317** | **0.9620** | **0.7962** | 0.7681 | 0.7732 |

TABLE 5
Results for causal span extraction task [54] on the test sets of RECCON-DD and RECCON-IE. DD stands for RECCON-DD with IE for RECCON-IE. The baseline results are reprinted from the literature [54].

| | | **EMpos↑** | **F1pos↑** | **F1neg↑** | **F1↑** |
|---|---|---|---|---|---|
| DD | RoBERTa | 32.63 | 58.17 | 85.85 | 75.45 |
| | SpanBERT | 34.64 | 60 | 86.02 | 75.71 |
| | HCL-ECPE | **38.12** | **63.40** | **88.71** | **78.8** |
| IE | RoBERTa | 10.19 | 26.88 | 91.68 | 84.52 |
| | SpanBERT | 22.41 | 37.8 | 90.54 | 82.86 |
| | HCL-ECPE | **25.35** | **44.20** | **92.7** | **84.81** |

ECPED has 2,154 emotion-cause pairs, 2,085 emotion clauses, and 2,142 cause clauses. RECCON is an emotion cause in the conversation dataset, which contains 1,122 dialogues and 11,769 utterances. RECCON consists of two parts, RECCON-IE and RECCON-DD. RECCON-IE denotes the samples are collected from IEMOCAP [55], and RECCON-DD denotes the samples are collected from DailyDialog [56]. The detailed statistics of datasets ECPED and RECCON are shown in Table 1 and Table 2, respectively.

TABLE 6
Results for causal emotion entailment task [54] on the test sets of RECCON-DD and RECCON-IE.

| | | Pos.F1↑ | Neg.F1↑ | macro F1↑ |
|---|---|---|---|---|
| DD | RoBERTa(Base) | 64.28 | 88.74 | 76.51 |
| | RoBERTa(Large) | 66.25 | 87.89 | **77.06** |
| | ECPE-MLL | 48.48 | 94.68 | 71.59 |
| | ECPE-2D | 55.50 | 94.96 | 75.23 |
| | RankCP | 33.00 | 97.30 | 65.15 |
| | HCL-ECPE | 36.43 | **98.26** | 67.86 |
| | HCL-ECPE(Base) | **66.47** | 88.52 | 76.93 |
| IE | RoBERTa(Base) | 28.02 | 95.67 | 61.85 |
| | RoBERTa(Large) | **40.83** | 95.68 | **68.26** |
| | ECPE-MLL | 20.23 | 93.55 | 57.65 |
| | ECPE-2D | 28.67 | 97.39 | 63.03 |
| | RankCP | 15.12 | 92.24 | 54.75 |
| | HCL-ECPE | 30.11 | **97.42** | 64.35 |
| | HCL-ECPE(Base) | 38.82 | 89.68 | 64.04 |

TABLE 7
Ablation study.

| | P | R | F |
|---|---|---|---|
| HCL-ECPE | **0.6569** | **0.6534** | **0.6537** |
| w/o $\mathcal{L}_{ICCL}$ | 0.6449 | 0.6271 | 0.6372 |
| w/o $\mathcal{L}_{IPCL}$ | 0.6522 | 0.6371 | 0.6467 |
| w/o $\mathcal{L}^e_{IPCL}$ | 0.6569 | 0.6405 | 0.6514 |
| w/o $\mathcal{L}^c_{IPCL}$ | 0.6577 | 0.6421 | 0.6532 |
| w/o $\{\mathcal{L}_{ICCL}, \mathcal{L}_{IPCL}\}$ | 0.6244 | 0.6306 | 0.6256 |

### 4.2 Implementation Detail

We use 200-dimension pre-trained Word2Vec [57] to initialize the word embedding. We also implement HCL-ECPE with the pre-trained BERT encoder [52] that is initialized using BERT-Base, Chinese [3]. The batch size and the dimension of clause representation are set to 32 and 200, respectively. For dataset ECPED, we set the dropout rates of 0.1, 0.5, 0.1, and 0.1 for the embedding layer, word-level Bi-LSTM, clause-level Bi-LSTM, and prediction layer, respectively. The hyperparameters $\{\alpha, \beta, \beta^e, \beta^c\}$ are set with 0.5, 0.3, 0.5 and 0.5, respectively. For dataset RECCON, we set the dropout rates of 0.2, 0.5, 0.1, and 0.1 for the embedding layer, word-level Bi-LSTM, clause-level Bi-LSTM, and prediction layer, respectively. The hyperparameters $\{\alpha, \beta, \beta^e, \beta^c\}$ are set with 0.5, 0.1, 0.5 and 0.5, respectively.

### 4.3 Evaluation Metrics

We use 10-fold cross-validation to evaluate model performance and conduct a one-sample t-test. In order to evaluate the performance of emotion clause extraction and cause clause extraction, we decompose the extracted emotion-cause pairs into the emotion clause set and cause clause set, and adopt the precision (P), recall (R), and F1 score (F1) as the evaluation metrics, which are defined as:

$$P = \frac{n^c}{N^c}, R = \frac{n_c}{M^c}, F1 = \frac{2 \times P \times R}{P + R}, \tag{20}$$

where $n_c$, $N^c$ and $M^c$ denote the number of correctly predicted results, detected results and real results, respectively. As [54] did, we use $\text{EM}_{\text{pos}}$, $\text{F1}_{\text{pos}}$, $\text{F1}_{\text{neg}}$, and F1 as metrics to evaluate the model performance on RECCON. $\text{EM}_{\text{pos}}$ represents the number of causal spans extracted by the model with respect to the gold standard data. $\text{F1}_{\text{pos}}$ aims to evaluate predictions of extractive QA models [58]. $\text{F1}_{\text{neg}}$ represents the F1 score of detecting negative examples with respect to the gold standard data.

### 4.4 Baseline Methods

We compare the existing works with HCL-ECPE, and the details of baseline methods are given as below:

- **Inter-EC** [11] firstly extracts the emotion and cause clauses, and then extracts emotion-cause pairs from the clause pairs composed by extracted emotion and cause clauses.
- **E2EECPE** [13] views ECPE as a link prediction from emotion clause to cause clause, and a link exists between them if the two clauses construct an emotion-cause pair.
- **PairGCN** [19] takes the candidate emotion-cause pairs as graph nodes and performs graph convolutional networks [59] to capture the dependency of the candidate pair in the local neighborhood.
- **LAE-MANN** [60] proposes **LML** and **LMB**, where **LML** takes LSTM as backbone and **LMB** adapts BERT as backbone.
- **ECPE-2D** [20] extends original Transformer to 2D-Transformer for modeling the mutual impacts across candidate emotion-cause pairs.
- **SLSN-U** [17] proposes a local search to extract the emotion clause and its cause clause simultaneously, and extracts the cause clause and its emotion clause in a similar way.
- **RANKCP** [14] utilizes graph attention networks (GATs) [61] to capture inter-clause dependency, and incorporates kernel-based relative position embedding to improve pair representation learning.
- **ECPE-MLL** [4] extracts emotion-cause pair extraction from dual perspectives through extracting the cause clause specified to the emotion clause and extracting the emotion clause specified to the cause clause.
- **Transition** [5] converts a sequence of actions to the directed graph with labeled edges and extracts emotion-cause pair through a procedure of parsing-like directed graph construction.
- **Tagging** [12] is a sequence labeling framework by incorporating the relative distance between the emotion clause and cause clause as a part of labeling tags to extract emotion-cause pairs.
- **IE-CNN-CRF** [3] regards ECPE as a sequence labeling problem, and considers the emotional type into the label to distinguish emotion cause under different emotion clauses and extract emotion-cause pairs.
- **UTOS** [18] designs a novel sequence label containing the content part and pairing part, where the content part extracts emotion and cause clause, and the pairing part is used to control the match between emotion clause and cause clause, so as to solve their extraction and pairing.
- **MGSAG** [62] builds a multi-granularity semantic aware graph model.

3. https://github.com/google-research/bert

- **EPO-ECPE** [63] proposes an end-to-end emotion-oriented emotion-cause pair extraction, aiming to fully exploit the benefits among emotion prediction, ground pair prediction and fake pair prediction.
- **UECA-Prompt** [64] proposes a universal prompt tuning method to solve different ECA tasks in the unified framework.
- **A2Net** [65] proposes a cross-task training method to further explore the model's capability.
- **Refinement** [35] regards ECPE as a sequence labeling problem and uses the prediction distribution of auxiliary tasks as an inductive bias to adjust the pair tagging distribution.
- **RoBERTa** [66] is a Transformer-based model, in which it takes [CLS] token and the emotion label [Et] as the front, and joins utterance pair with [SEP] in between to create the input.
- **SpanBERT** [67] is a pre-trained model following a training objective, i.e., predicting masked contiguous spans instead of tokens.

TABLE 8
The seven documents for case study.

| DID | Document Length | Emotion-Cause Pair |
|---|---|---|
| 72 | 14 | (7,5),(7,6) |
| 82 | 8 | (8,7) |
| 127 | 14 | (4,2) |
| 207 | 19 | (14,17) |
| 387 | 13 | (7,6), (11,10) |
| 400 | 10 | (5, 4), (9, 7), (9, 8) |
| 455 | 18 | (3,2),(14,13) |

### 4.5 Results on Emotion-Cause Pair Extraction

Table 3 and Table 4 report the results without and with BERT as encoder, respectively. In Table 3, Inter-EC's two-step setting leads to low recalls on emotion-cause pair extraction, emotion extraction, and cause extraction. E2EECPE adopts a unified architecture to solve the error propagation caused by the two-step setting and significantly improves the recall rate. Meanwhile, the precision of E2EECPE on the emotion-cause pair extraction decreases. PairGCN makes a better precision but loses a degree of recall on emotion-cause pair extraction. ECPE-MLL significantly improves precision, but its recall slightly decreases. We can observe that HCL-ECPE outperforms most of the competitive baselines, and its F1 score surpasses Inter-EC by 4.09%, E2EECPE by 2.57%, and PairGCN by 2.16%, respectively. The improvements demonstrate the effectiveness of HCL-ECPE in emotion-cause pair extraction. To verify the orthogonality of ICCL and IPCL, we introduce ICCL and IPCL as the regularization terms of E2EECPE [13], RANKCP [14], ECPE -MLL [4], UTOS [18], and EPO-ECPE [63]. It can be observed that the model's F1 scores are improved by 3.52%, 1.72%, 1.16%, 2.57%, and 2.32% for emotion-cause pair extraction, respectively. These results demonstrate that ICCL and IPCL modules are orthogonal, and introducing them can bring significant improvements for the existing model.

For Table 4, HCL-ECPE still outperforms most of the competitive baselines, including competitive models like PairGCN using GCN [59] and ECPE-2D using 2D-Transformer [34]. HCL-ECPE with the BERT as an encoder can obtain improvements on all evaluation metrics in three tasks, which illustrates the strong semantic encoding capability of BERT. We speculate that the reasons that HCL-ECPE fails to outperform most baselines can be summarized in two folds. Note that most baselines in Table 3 and Table 4 adapt the strong encoders. For example, RANKCP uses graph attention networks [61] to encode the clause representations, and ECPE-2D employs multiple Transformer [34] layers to model inter-clause dependencies. EPO-ECPE explores the benefits of emotion prediction to improve the performance of emotion-cause pair extraction. Compared with these baselines, HCL-ECPE adopts a simple encoder, i.e., hierarchical encoder, as the backbone to regularize the clause representation and pair representation. Additionally, most baselines set context windows based on the characteristics that the emotion clause and its cause clause are close to each other, thereby reducing extraction difficulty. For example, ECPE-MLL and SLSN-U detect the cause clause of the specified emotion clause or the emotion clause of the specified cause clause in a local range. That is, some works use the relative position between the emotion clause and cause as the indication to model the relation between them. However, the model's performance is vulnerable to the changes of relative position [66], which damages the robustness of the model to some extent [24], [25]. In contrast with these methods, HCL-ECPE does not utilize the position feature preliminarily to improve model performance. Furthermore, we introduce ICCL and IPCL modules as the contrastive learning loss of RANKCP, ECPE-MLL, UTOS, A2Net, and EPO-ECPE. It is observed that introducing ICCL and IPCL modules can further improve model performance by 1.52%, 1.14%, 1.6%, 0.58%, and 1.92% in the F1 score, respectively.

Meanwhile, we introduce more strong networks like GCN [59], Transformer [34] as the encoders to model inter-clause dependency. We can observe that stronger encoders can bring performance improvements. Compared with HCL-ECPE, HCL-ECPE(GCN) and HCL-ECPE(Transformer) obtain 2.03% and 3% in F1 score of emotion-cause pair extraction, respectively. We also conduct a one-sample t-test on the F1 scores in the three tasks, and the improvements are all statistically significant with $p < 0.05$. These results illustrate the effectiveness of HCL-ECPE in representation learning and further verify the orthogonality of ICCL and IPCL to the existing models.

### 4.6 Results on Causal Span Extraction and Causal Emotion Entailment tasks

Causal span extraction is a task that aims to identify the causal emotion-cause span for the non-neutral utterance [54]. Table 5 shows the results of the causal span extraction task for RECCON-DD and RECCON-IE datasets. HCL-ECPE achieves the best performance whether it performs on RECCON-DD or RECCON-IE. As with most ECPE works, HCL-ECPE extracts the causal span with contextual information from the conversation. HCL-ECPE obtains noticeable margin improvements compared with RoBERTa and SpanBERT. For RECCON-DD, HCL-ECPE outperforms SpanBERT by nearly 3.48%, 3.4%, 2.69%, and 3.09% in metrics

TABLE 9
Comparison of predicted emotion-cause pairs between RANKCP, RANKCP(HCL) and HCL-ECPE.

| DID | Emotion-cause Pair | RANKCP Prediction | RANKCP(HCL) Prediction | HCL-ECPE Prediction |
|---|---|---|---|---|
| 82 | ($c_7$) Let children grow up healthier and more sunny, ($c_8$) and I hope to express my gratitude to Houma Public Security Bureau, Houma volunteer team and all kind-hearted people who care about helping their family find xiaoxinxin through this newspaper | $(c_8, c_7)$ ✓ | $(c_8, c_7)$ ✓ | $(c_8, c_7)$ ✓ |
| 127 | ($c_2$) mentioning an Jirong's support to his father over the years, ($c_3$) at noon on January 3, ($c_4$) Guo Chunying, 50, Guo Tianlu's second daughter, had red eyes | $(c_4, c_3)$ ✗ | $(c_4, c_2)$ ✓ | $(c_4, c_2)$ ✓ |
| 72 | ($c_5$)But I can't spend the new year with my children, ($c_6$) and the parents are getting older and older, ($c_7$) I always feel guilty | $(c_7, c_6)$ ✗ | $(c_7, c_5), (c_7, c_6)$ ✓ | $(c_7, c_5), (c_7, c_6)$ ✓ |
| 207 | ($c_{14}$) Yu Jingmin was a little worried about the reporter's interview, ($c_{15}$) he said, ($c_{16}$) my whole life has passed, ($c_{17}$) don't want others to disturb | $(c_{14}, c_{15})$ ✗ | $(c_{14}, c_{17})$ ✓ | $(c_{14}, c_{17})$ ✓ |
| 387 | ($c_6$) A businessman is unlikely to turn off his mobile phone, ($c_7$) Li Min suddenly became suspicious, ($c_{10}$) She found several ambiguous messages in WeChat, ($c_{11}$) couldn't help getting angry | $(c_7, c_6)$ ✗ | $(c_7, c_6), (c_{11}, c_{10})$ ✓ | $(c_7, c_6), (c_{11}, c_{10})$ ✓ |
| 400 | ($c_4$) These ordinary words, ($c_5$) moved everyone on the scene. ($c_7$) he learned that my daughter has been married, ($c_8$) and given birth to a child, ($c_9$) Kong Qinghe is very excited | $(c_9, c_8)$ ✗ | $(c_5, c_4), (c_9, c_7), (c_9, c_8)$ ✓ | $(c_9, c_7), (c_9, c_8)$ ✗ |
| 455 | ($c_2$) Miss Fan remembers falling into the water, ($c_3$) Still have lingering fears. ($c_{13}$) he didn't expect to slide into the river from the grass, ($c_{14}$) and he was very nervous | $(c_{14}, c_{13})$ ✗ | $(c_3, c_2), (c_{14}, c_{13})$ ✓ | $(c_{14}, c_{13})$ ✗ |

EMpos, F1pos, F1neg, and F1, respectively. Similarly, HCL-ECPE reaps consistent improvements on all metrics for the RECCON-IE dataset. These improvements illustrate the capacity of HCL-ECPE on causal span extraction, thereby indicating that HCL-ECPE can solve span-level emotion-cause extraction problems. Furthermore, we also perform HCL-ECPE on the causal emotion entailment task, which can be regarded as a simpler version of the span extraction task [54]. The results are shown in Table 6. This task aims to predict which particular utterances in the conversation history are responsible for the given target non-neutral utterance. It can be observed that HCL-ECPE outperforms the baselines ECPE-MLL, ECPE-2D, and RANKCP, on all evaluation metrics for RECCON-DD and RECCON-IE datasets. However, compared to RoBERTa [66], there is still a significant gap between HCL-ECPE and RoBERTa(Base) or RoBERTa(Large). To make a fair comparison, we take the encoder of RoBERTa to replace the hierarchical encoder of HCL-ECPE to extract the emotion-cause pairs from a document. It can be observed that HCL-ECPE(Base) slightly outperforms RoBERTa(Base) and is close to RoBERTa(Large) in model performance in both RECCON-DD and RECCON-IE datasets.

### 4.7 Re-evaluating ECPE task

Following the works of [24], we reproduce the results of typical models after removing the relative distance embedding, and the results are shown in Table 10. We can observe that removing the relative distance embedding damages the performance of Inter-EC, Inter-CE, ECPE-2D, RANKCP, and EPO-ECPE, and their F1 scores of emotion-cause pair extraction drop around 2%, 5.63%, 2.98%, 3.5% and 7%, respectively. These models are position-sensitive and the performances are easily affected by the relative position changing between the emotion and cause clauses [24]. Compared with these models, our proposed model is less sensitive to the relative distance because it depends on the statistical relationship between the emotion clause and cause clause instead of their position relation, so it shows a more stable performance. These results demonstrate that extracting emotion-cause pairs from the perspective of mutual information is effective and can improve the model's robustness.

Additionally, [24] reconstructed a new "de-bias" dataset for the ECPE task based on the original ECE de-bias dataset [25]. The "de-bias" dataset balances the distribution of relative distances between the emotion clauses and the cause clauses, aiming to alleviate the bias of benchmark dataset. We conduct experiments on the reconstructed "de-bias" dataset, and the results are shown in Table 11. It can be observed that the F1 scores of some competitive models like Inter-EC, ECPE-2D, RANKCP, ECPE-MLL, and EPO-ECPE all decrease significantly on the "de-bias"dataset, which indicates again that these models are sensitive to the changing of position relation between emotion clause and cause clause, and have weak model's robustness in emotion-cause pair extraction. Compared with the best performance, our proposed model performs better and obtains 4.85% improvements in the F1 score on the "de-bias" dataset. These results demonstrate that HCL-ECPE can mine the intrinsic dependencies between the emotion clause and cause clause by measuring their mutual information, rather than the position dependencies between the emotion clause and cause clause.

TABLE 10
Comparison results after removing the relative distance embeddings.

| | P | R | F1 |
|---|---|---|---|
| Inter-EC w/o RP | 0.5983 | 0.5497 | 0.5701 |
| Inter-CE w/o RP | 0.5737 | 0.545 | 0.5565 |
| ECPE-2D w/o RP | 0.6711 | 0.5758 | 0.6198 |
| RANKCP w/o RP | 0.6378 | 0.6160 | 0.6260 |
| ECPE-MLL w/o RP | 0.5511 | 0.6617 | 0.6010 |
| EPO-ECPE w/o RP | 0.5392 | 0.6917 | 0.6140 |
| HCL-ECPE | **0.6569** | **0.6534** | **0.6537** |

### 4.8 Ablation Study

The benchmark dataset of ECPE exhibits position bias, where the phenomenon of position bias refers to the tendency for cause clauses in most samples to appear near to their corresponding emotion clauses. We reproduce the results of typical baselines like Inter-EC, ECPE-2D, RANKCP, and ECPE-MLL after eliminating relative distance embeddings, and the results are given in Table 7. We find that removing relative distance embedding from the baselines Inter-EC, ECPE-2D, and RANKCP leads to the drops of F1 scores by 2.1%, 3.5%, and 7.3%, respectively. Compared with these models, HCL-ECPE achieves 65.37% in F1 score on emotion-cause pair extraction, outperforming the existing models in the case of removing relative position. This result indicates that the HCL-ECPE shows effective representation learning and exhibits more robust emotion-cause pair extraction. We also conduct an ablation study on HCL-ECPE to verify the

TABLE 11
The performances of HCL-ECPE as well as other models on "de-bias" dataset.

| | P | R | F1 |
|---|---|---|---|
| Inter-EC | 0.4302 | 0.2548 | 0.3171 |
| ECPE-2D | 0.4722 | 0.3738 | 0.4173 |
| RANKCP | 0.4368 | 0.3916 | 0.4122 |
| EPO-ECPE | 0.3998 | 0.4476 | 0.4248 |
| HCL-ECPE | **0.5065** | **0.4514** | **0.4733** |

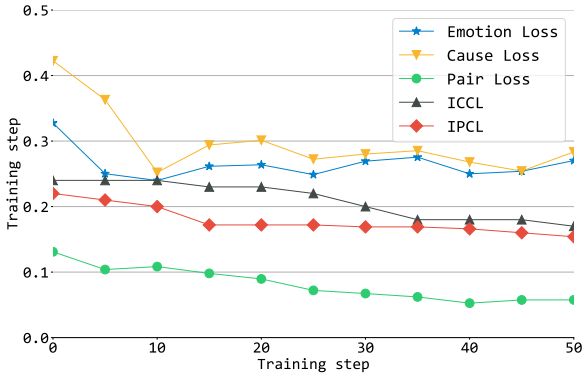



Fig. 2. The illustration of loss curves in the training phase, including the losses of emotion extraction, cause extraction, emotion-cause pair extraction, ICCL and IPCL.

effectiveness of contrastive learning loss by eliminating one or several loss terms from the total loss function. First, we eliminate $\mathcal{L}_{ICCL}$, which leads to a performance degradation of about 1.65% in the F1 score on emotion-cause pair extraction, indicating that the ICCL module can improve model performance. Then we remove module $\mathcal{L}_{IPCL}$ from HCL-ECPE, resulting in a decrease of the F1 score by 0.6% on emotion-cause pair extraction, which verifies the influential role of the IPCL module. More specifically, we individually remove $\mathcal{L}^e_{IPCL}$ and $\mathcal{L}^c_{IPCL}$ from HCL-ECPE to analyze the effects of the emotion clause on the emotion-cause pair and cause clause on the emotion-cause pair, respectively. It can be observed that the F1 scores after removing $\mathcal{L}^e_{IPCL}$ and $\mathcal{L}^c_{IPCL}$ decrease by 0.23% and 0.05% respectively. These results again demonstrate the effectiveness of intra-pair contrastive learning in improving clause pair representation.

In addition,we tracked loss changing in different stages of training, as shown in Figure 2. It is observed that the loss curves of emotion extraction and cause extraction are steeper than the curves of ICCL and IPCL modules, indicating that the losses of emotion extraction and cause extraction rapidly decrease after the first few training steps. Furthermore, it can be observed that the losses of ICCL and IPCL modules are smaller than the losses of emotion extraction and cause extraction, which indicates that ICCL and IPCL as the regualizer terms do not bring the disturbance to the training of emotion extraction and cause extraction.

### 4.9 Case Study

We randomly select some samples to conduct the case study. Table 8 gives the information on document ID (DID), document length, and emotion-cause pairs for seven representative cases. These seven cases cover single emotion-cause pair documents, i.e., document 82; small RP documents, i.e., document 127; large RP documents, i.e., document 207; and multiple emotion-cause pair documents, i.e., documents 72, 387, 400, and 455. For the multiple emotion-cause pair documents, the emotion-cause pairs in document 72 have the same emotion clause, while the emotion-cause pairs in documents 387, 400 and 455 have different emotion clauses. We give the HCL-ECPE's predictions on these documents and compare them with the predictions of RANKCP and RANKCP(HCL). The results are shown in Table 9. Document 82 has single emotion-cause pair, where the cause clause $c_7$ is close to its emotion clause $c_8$. We can observe that RANKCP, RANKCP(HCL), and HCL-ECPE can correctly predict the emotion-cause pair $(c_8, c_7)$. For document 127, HCL-ECPE and RANKCP(HCL) correctly predict the emotion-cause pair $(c_4, c_2)$, and RANKCP extracts the wrong pair $(c_4, c_3)$ as the emotion-cause pair. The reason for the wrong prediction is that the kernel-based relative position embedding setting of RANKCP tends to predict those clauses that are close to the emotion clause as its cause clauses. Compared with RANKCP, hierarchical contrastive learning focuses on the relative position between emotion and cause in the feature space instead of their relative position in the document.

A crucial challenge of the ECPE task is to extract multiple emotion-cause pairs within a document. Most existing models only extract one of the emotion-cause pairs for multiple emotion-cause pair documents. To verify the effectiveness of hierarchical contrastive learning modules on multiple emotion-cause pair extraction, we analyze the predictions of RANKCP(HCL) and HCL-ECPE for multiple emotion-cause pair documents, and compare them with RANKCP. For document 72, HCL-ECPE and RANKCP(HCL) can extract all emotion-cause pairs simultaneously while RANKCP only extracts $(c_7, c_6)$. The reason is that the hierarchical contrastive learning module pulls the representations of multiple emotion clauses and cause clauses in the same document close in embedding space, thereby helping provide more information to determine the possible emotion-cause pairs. Document 400 has three emotion-cause pairs, of which two emotion-cause pairs, i.e., $(c_9, c_7)$ and $(c_9, c_8)$, have the same emotion clause, and the remaining $(c_5, c_4)$ has a different emotion clause. In this case, RANKCP and HCL-ECPE merely predict $(c_9, c_8)$ as the emotion-cause pairs, while RANKCP(HCL) can predict all the emotion-cause pairs. This phenomenon also appears in document 455. RANKCP and HCL-ECPE predict a part of emotion-cause pairs, while RANKCP(HCL) predicts all emotion-cause pairs. This may be because hierarchical contrastive learning modules recognize emotions and their causes by detecting their position relation in embedding space, which is conducive to finding multiple emotions simultaneously. Compared with HCL-ECPE, RANKCP(HCL) has a more powerful component to learn clause representation, such as a GNN-based encoder and kernel-based position embedding, which maybe the reason that RANKCP(HCL) is superior than HCL-ECPE. Moreover, when the cause clause is far from the emotion clause, as shown in document 207, HCL-ECPE and RANKCP(HCL) still can correctly extract emotion-cause pairs, but RANKCP fails to extract them. These results also demonstrate the superiority of the hierarchical contrastive

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2024.3391854



TABLE 12
Examples of the predicted emotion-cause pairs, where the first column depicts the content of the emotion-cause pair, the second column gives the ground-truth label, and the third column depicts the prediction of HCL-ECPE. The emotion clause is denoted with $c^*$, the cause clause is denoted with $c^{\S}$, and the correct prediction is marked with underline.

| Emotion-cause Content | Ground Label | Prediction Label |
|---|---|---|
| $[\cdots]$, $(c_2^{\S})$ he has not married after thirty, $(c_3^*)$ the family is very anxious, $(c_4)$ repeated blind dates and repeated rejections, $[\cdots]$, $(c_{10}^{*\S})$ Wang was very moved because Chen didn't dislike his appearance, $[\cdots]$ | $(c_3, c_2)$, $\underline{(c_{10}, c_{10})}$ | $\underline{(c_{10}, c_{10})}$ |
| $[\cdots]$, $(c_3^{\S})$ there is a woman standing on the electric tower tottering, $(c_4^*)$The villagers and passers-by were extremely anxious, $[\cdots]$, $(c_8)$ a firefighter slowly climbed to a place close to her 10 meters, $(c_9)$ soothe her mood, $[\cdots]$ | $(c_4, c_3)$ | $(c_9, c_8)$ |
| $(c_1)$ According to Radio Hong Kong, $[\cdots]$, $(c_{14})$ judge described, $(c_{15}^*)$the case is a sad tragedy, $(c_{16}^{\S})$ the stabbed defendant his wife with a sharp knife so hard that his wife's sternum was broken and her main organs were seriously damaged, $[\cdots]$ | $(c_{15}, c_{16})$ | $(c_{16}, c_{14})$,$(c_{16}, c_{15})$ |
| $[\cdots]$, $(c_2^{\S})$ It is found that the woman's appearance is very different from the photos and her body shape is also very different, $[\cdots]$, $(c_7^*)$ he is greatly disappointed, $(c_8)$ The man thought he had been deceived, $[\cdots]$ | $\underline{(c_7, c_2)}$ | $\underline{(c_7, c_8)}$ |

learning modules.

### 4.10 Error Analysis

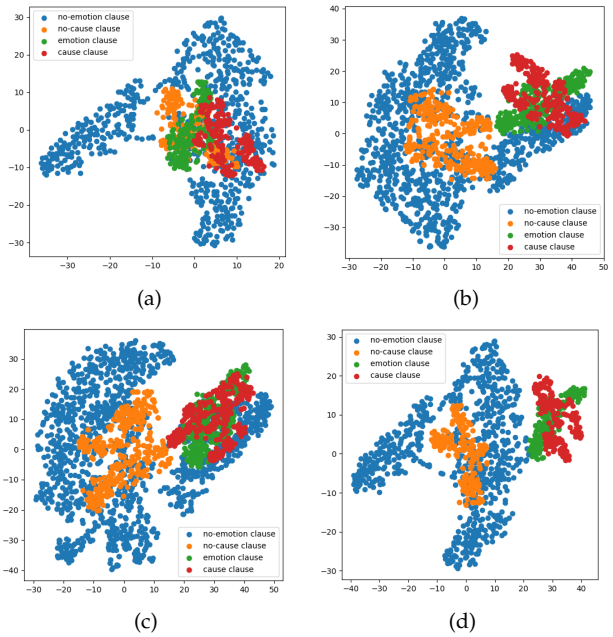
To further understand the proposed model HCL-ECPE, we conduct the error analysis to analyze the prediction results of HCL-ECPE. The results are shown in Table 12. We summarize the prediction errors into three categories, i.e., incomplete prediction error, wrong emotion prediction, and wrong cause prediction.

The first is the incomplete prediction error. When a document has two or more emotion-cause pairs, the model predicts one of the emotion-cause pairs and omits the other emotion-cause pairs, which results in an incomplete prediction. For example, HCL-ECPE correctly predicts the emotion-cause pairs $(c_{10}, c_{10})$ in the first document. However, HCL-ECPE neglects the emotion-cause pair $(c_3, c_2)$. The second is the wrong emotion prediction. For the second document, we can observe that HCL-ECPE incorrectly predicts $c_9$ as an emotion clause and makes a wrong prediction on the cause clause. Generally, the model cannot detect correctly cause clause correctly once it predicts the wrong emotion clause, since the model detects the cause clause based on the predicted emotion clause. Similarly, in the third document, HCL-ECPE makes the wrong prediction on the emotion-cause pair due to the wrong emotion prediction. The last error is wrong cause prediction. In the fourth document, HCL-ECPE correctly predicts the emotion clause $c_7$ but makes a wrong pairing caused by the wrong cause clause.

### 4.11 Visualization

To examine the effects of ICCL and IPCL modules on representation learning, we store the representations of test samples and visualize them using the T-SNE tool. We extract the clause representations in the four cases: 1) emotion clause $r_i^e, i \in E$, 2) cause clause $r_j^c, j \in C$, 3) non-emotion clause $r_i^e, i \in \bar{E}$ and 4) non-cause clause $r_j^c, j \in \bar{C}$. $E$, $C$, $\bar{E}$, and $\bar{C}$ denote the emotion clause set, cause clause set, non-emotion clause set, and non-cause clause set, respectively. We visualize clause representation in the following four cases: 1) the representation learned after removing hierarchical contrastive learning, shown in Figure 3(a), 2) the representation learned after removing the ICCL module, shown in Figure 3(b), 3) the representation learned after removing IPCL module, shown in Figure 3(c), and 4) the representation learned with the hierarchical contrastive learning, shown in Figure 3(d). We can observe from Figure 3(a) that the representations of emotion and cause clauses share some data points with the non-emotion and non-cause clauses in their low-dimension representation, which brings disturbance for the model to distinguish emotion/cause from non-emotion/non-cause. Compared with the representation without contrastive learning, incorporating modules ICCL or IPCL enables emotion/cause far from the non-emotion/non-cause in representation space, as shown in Figure 3(b) and Figure 3(c). Furthermore, when the proposed model considers the effects of ICCL and IPCL simultaneously, it is

Fig. 3. T-SNE visualization comparisons of the emotion, cause, non-emotion, and non-cause clause representations learned by HCL-ECPE (a) without hierarchical contrastive learning, (b) without ICCL module, (c) without IPCL module, and (d) with hierarchical contrastive learning.



more distinguishable between the representations of emotion/cause and the non-emotion/non-cause clause, as shown in Figure 3(d). These results indicate that the proposed ICCL and IPCL can effectively help the model extract emotion and cause clauses from clauses. Moreover, the visualizations also demonstrate that the representations of emotion and cause clauses are pulled close with the help of a hierarchical contrastive learning module, thereby capturing the pairing relation between them in the representation space.

## 5 Conclusion

This paper proposes a hierarchical contrastive learning framework (HCL-ECPE) for solving ECPE, which improves representation learning with inter-clause and intra-pair contrastive learning. Inter-clause contrastive learning (ICCL) pulls close the emotion clause representation and cause clause representation through mutual information maximization. Intra-pair contrastive learning (IPCL) learns the encoding path between emotion/cause clause representation and emotion-cause pair representation to implement joint learning. We conduct a series of experiments, and the results demonstrate that HCL-ECPE outperforms most competitive baselines. Furthermore, experimental results demonstrate that the proposed ICCL and IPCL modules are orthogonal to the existing works, and it is believed that introducing ICCL and IPCL can bring improvements.

## References


[1] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "Unimse: Towards unified multimodal sentiment analysis and emotion recognition," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 7837–7851.

[2] Z. Zhu, X. Cheng, Y. Li, H. Li, and Y. Zou, "Aligner$^2$: Enhancing joint multiple intent detection and slot filling via adjustive and forced cross-task alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 777–19 785.

[3] X. Chen, Q. Li, and J. Wang, "A unified sequence labeling model for emotion cause pair extraction," in *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, 2020, pp. 208–218. [Online]. Available: https://doi.org/10.18653/v1/2020.coling-main.18

[4] Z. Ding, R. Xia, and J. Yu, "End-to-end emotion-cause pair extraction based on sliding window multi-label learning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 2020, pp. 3574–3583. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.290

[5] C. Fan, C. Yuan, J. Du, L. Gui, M. Yang, and R. Xu, "Transition-based directed graph construction for emotion-cause pair extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 2020, pp. 3707–3717. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.342

[6] G. Hu, G. Lu, and Y. Zhao, "Emotion-cause joint detection: A unified network with dual interaction for emotion cause analysis," in *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9*. Springer, 2020, pp. 568–579.

[7] S. Y. M. Lee, Y. Chen, and C.-R. Huang, "A text-driven rule-based system for emotion cause detection," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles, CA: Association for Computational Linguistics, Jun. 2010, pp. 45–53. [Online]. Available: https://www.aclweb.org/anthology/W10-0206

[8] L. Gui, D. Wu, R. Xu, Q. Lu, and Y. Zhou, "Event-driven emotion cause extraction with corpus construction," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016, pp. 1639–1649. [Online]. Available: https://www.aclweb.org/anthology/D16-1170/

[9] X. Li, K. Song, S. Feng, D. Wang, and Y. Zhang, "A co-attention neural network model for emotion cause analysis with emotional context awareness," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 2018, pp. 4752–4757. [Online]. Available: https://www.aclweb.org/anthology/D18-1506/

[10] H. Yan, L. Gui, G. Pergola, and Y. He, "Position bias mitigation: A knowledge-aware graph model for emotion cause extraction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 2021, pp. 3364–3375. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.261

[11] R. Xia and Z. Ding, "Emotion-cause pair extraction: A new task to emotion analysis in texts," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, 2019, pp. 1003–1012.

[12] C. Yuan, C. Fan, J. Bao, and R. Xu, "Emotion-cause pair extraction as sequence labeling based on A novel tagging scheme," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 2020, pp. 3568–3573. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.289

[13] H. Song, C. Zhang, Q. Li, and D. Song, "End-to-end emotion-cause pair extraction via learning to link," *CoRR*, vol. abs/2002.10710, 2020. [Online]. Available: https://arxiv.org/abs/2002.10710

[14] P. Wei, J. Zhao, and W. Mao, "Effective inter-clause modeling for end-to-end emotion-cause pair extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 2020, pp. 3171–3181. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.289

[15] G. Hu, G. Lu, and Y. Zhao, "Bidirectional hierarchical attention networks based on document-level context for emotion cause extraction," in *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, 2021, pp. 558–568. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-emnlp.51

[16] ——, "FSS-GCN: A graph convolutional networks with fusion of semantic and structure for emotion cause analysis," *Knowl. Based Syst.*, vol. 212, p. 106584, 2021. [Online]. Available: https://doi.org/10.1016/j.knosys.2020.106584

[17] Z. Cheng, Z. Jiang, Y. Yin, H. Yu, and Q. Gu, "A symmetric local search network for emotion-cause pair extraction," in *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, 2020, pp. 139–149. [Online]. Available: https://doi.org/10.18653/v1/2020.coling-main.12

[18] Z. Cheng, Z. Jiang, Y. Yin, N. Li, and Q. Gu, "A unified target-oriented sequence-to-sequence model for emotion-cause pair extraction," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2779–2791, 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3102194

[19] Y. Chen, W. Hou, S. Li, C. Wu, and X. Zhang, "End-to-end emotion-cause pair extraction with graph convolutional network," in *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, 2020, pp. 198–207. [Online]. Available: https://doi.org/10.18653/v1/2020.coling-main.17

[20] Z. Ding, R. Xia, and J. Yu, "ECPE-2D: emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 2020, pp. 3161–3170. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.288

[21] P. Lin, M. Yang, and Y. Gu, "A hierarchical inter-clause interaction network for emotion cause extraction," in *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, 2021, pp. 1–8. [Online]. Available: https://doi.org/10.1109/IJCNN52387.2021.9534291

[22] W. Yu and C. Shi, "Emotion cause extraction by combining

intra-clause sentiment-enhanced attention and inter-clause consistency interaction," in *6th IEEE International Conference on Computer and Communication Systems, ICCCS 2021, Chengdu, China, April 23-26, 2021*, 2021, pp. 146–150. [Online]. Available: https://doi.org/10.1109/ICCCS52626.2021.9449281

[23] R. Xia, M. Zhang, and Z. Ding, "RTHN: A rnn-transformer hierarchical network for emotion cause extraction," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 2019, pp. 5285–5291. [Online]. Available: https://doi.org/10.24963/ijcai.2019/734

[24] G. Hu, Y. Zhao, G. Lu, F. Yin, and J. Chen, "An exploration of mutual information based on emotion–cause pair extraction," *Knowledge-Based Systems*, vol. 256, p. 109822, 2022.

[25] J. Ding and M. Kejriwal, "An experimental study of the effects of position bias on emotion causeextraction," *CoRR*, vol. abs/2007.15066, 2020. [Online]. Available: https://arxiv.org/abs/2007.15066

[26] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 2021, pp. 9180–9192. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.723

[27] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. [Online]. Available: https://openreview.net/forum?id=Bklr3j0cKX

[28] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, 2010, pp. 297–304. [Online]. Available: http://proceedings.mlr.press/v9/gutmann10a.html

[29] Y. Chen, S. Y. M. Lee, S. Li, and C. Huang, "Emotion cause detection with linguistic constructions," in *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, 2010, pp. 179–187. [Online]. Available: https://www.aclweb.org/anthology/C10-1021/

[30] I. Russo, T. Caselli, F. Rubino, E. Boldrini, and P. Martínez-Barco, "Emocause: An easy-adaptable approach to extract emotion cause contexts," in *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA@ACL 2011, Portland, OR, USA, June 24, 2011*, 2011, pp. 153–160. [Online]. Available: https://www.aclweb.org/anthology/W11-1720/

[31] C. Fan, H. Yan, J. Du, L. Gui, L. Bing, M. Yang, R. Xu, and R. Mao, "A knowledge regularized hierarchical approach for emotion cause analysis," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 2019, pp. 5613–5623. [Online]. Available: https://doi.org/10.18653/v1/D19-1563

[32] L. Gui, J. Hu, Y. He, R. Xu, Q. Lu, and J. Du, "A question answering approach to emotion cause extraction," *CoRR*, vol. abs/1708.05482, 2017. [Online]. Available: http://arxiv.org/abs/1708.05482

[33] Y. Chen, W. Hou, X. Cheng, and S. Li, "Joint learning for emotion classification and emotion cause detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 2018, pp. 646–651. [Online]. Available: https://www.aclweb.org/anthology/D18-1066/

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[35] C. Fan, C. Yuan, L. Gui, Y. Zhang, and R. Xu, "Multi-task sequence tagging for emotion-cause pair extraction via tag distribution refinement," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2339–2350, 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3089837

[36] Q. Sun, Y. Yin, and H. Yu, "A dual-questioning attention network for emotion-cause pair extraction with context awareness," in *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, 2021, pp. 1–8. [Online]. Available: https://doi.org/10.1109/IJCNN52387.2021.9533767

[37] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 2020, pp. 1597–1607. [Online]. Available: http://proceedings.mlr.press/v119/chen20j.html

[38] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12356. Springer, 2020, pp. 776–794. [Online]. Available: https://doi.org/10.1007/978-3-030-58621-8_45

[39] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 9726–9735. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00975

[40] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, April 26 - May 1, 2015*, 2015, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ITW.2015.7133169

[41] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. [Online]. Available: https://openreview.net/forum?id=HyxQzBceg

[42] Y. Liu, W. Rong, and Z. Xiong, "Improved text matching by enhancing mutual information," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018, pp. 5269–5276. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16214

[43] D. Qian and W. K. Cheung, "Enhancing variational autoencoders with mutual information neural estimation for text generation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 2019, pp. 4045–4055. [Online]. Available: https://doi.org/10.18653/v1/D19-1416

[44] P. Cheng, M. R. Min, D. Shen, C. Malon, Y. Zhang, Y. Li, and L. Carin, "Improving disentangled text representation learning with information-theoretic guidance," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 2020, pp. 7530–7541. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.673

[45] Y. Yeh and Y. Chen, "Qainfomax: Learning robust question answering system by mutual information maximization," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 2019, pp. 3368–3373. [Online]. Available: https://doi.org/10.18653/v1/D19-1333

[46] A. P. B. Veyseh, F. Dernoncourt, M. T. Thai, D. Dou, and T. H. Nguyen, "Multi-view consistency for relation extraction via mutual information and structure prediction," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 2020, pp. 9106–9113. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6445

[47] I. Belghazi, S. Rajeswar, A. Baratin, R. D. Hjelm, and A. C. Courville, "MINE: mutual information neural estimation," *CoRR*, vol. abs/1801.04062, 2018. [Online]. Available: http://arxiv.org/abs/1801.04062

[48] Q. Minping and M. Silverstein, "On donsker and varadhan's asymptotic evaluation without compactness," *Acta Mathematicae Applicatae Sinica*, vol. 1, no. 1, pp. 17–25, 1984.

[49] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning

with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748

[50] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[51] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification," in *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 2016, pp. 1480–1489. [Online]. Available: https://doi.org/10.18653/v1/n16-1174

[52] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[53] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, 2006, pp. 1735–1742. [Online]. Available: https://doi.org/10.1109/CVPR.2006.100

[54] S. Poria, N. Majumder, D. Hazarika, D. Ghosal, R. Bhardwaj, S. Y. B. Jian, P. Hong, R. Ghosh, A. Roy, N. Chhaya, A. F. Gelbukh, and R. Mihalcea, "Recognizing emotion cause in conversations," *Cogn. Comput.*, vol. 13, no. 5, pp. 1317–1332, 2021. [Online]. Available: https://doi.org/10.1007/s12559-021-09925-7

[55] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359, 2008. [Online]. Available: https://doi.org/10.1007/s10579-008-9076-6

[56] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, 2017, pp. 986–995. [Online]. Available: https://aclanthology.org/I17-1099/

[57] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 2013, pp. 3111–3119.

[58] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100, 000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2016, pp. 2383–2392. [Online]. Available: https://doi.org/10.18653/v1/d16-1264

[59] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[60] H. Tang, D. Ji, and Q. Zhou, "Joint multi-level attentional model for emotion detection and emotion-cause pair extraction," *Neurocomputing*, vol. 409, pp. 329–340, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2020.03.105

[61] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[62] Y. Bao, Q. Ma, L. Wei, W. Zhou, and S. Hu, "Multi-granularity semantic aware graph model for reducing position bias in emotion cause pair extraction," in *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, 2022, pp. 1203–1213. [Online]. Available: https://doi.org/10.18653/v1/2022.findings-acl.95

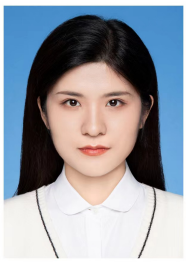
[63] G. Hu, Y. Zhao, and G. Lu, "Emotion prediction oriented method with multiple supervisions for emotion-cause pair extraction," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1141–1152, 2023. [Online]. Available: https://doi.org/10.1109/TASLP.2023.3250833

[64] X. Zheng, Z. Liu, Z. Zhang, Z. Wang, and J. Wang, "Ueca-prompt: Universal prompt for emotion cause analysis," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, 2022, pp. 7031–7041. [Online]. Available: https://aclanthology.org/2022.coling-1.613

[65] S. Chen, X. Shi, J. Li, S. Wu, H. Fei, F. Li, and D. Ji, "Joint alignment of multi-task feature and label spaces for emotion cause pair extraction," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, 2022, pp. 6955–6965. [Online]. Available: https://aclanthology.org/2022.coling-1.606

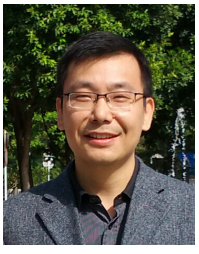
[66] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[67] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 64–77, 2020. [Online]. Available: https://doi.org/10.1162/tacl_a_00300



**Guimin Hu** received the B.S. degree in information management and information system from Liaocheng University, China, in 2015 and received the M.S. degree in computer science and technology from Northeastern University, China, in 2018. She received the Ph.D. degree from the Harbin Institute of Technology Shenzhen, China, in 2023. Now, she is a Postdoc at University of Copenhagen. Her current research interests include emotion cause analysis, multi-modal sentiment analysis, and emotion recognition in conversation.

**Yi Zhao** received the Ph.D. degree in electronic engineering (nonlinear dynamics) from Hong Kong Polytechnic University, Hong Kong, in 2007. Since 2007, he has been with the Harbin Institute of Technology, Shenzhen,China, where he is currently a Professor. His research interests include applied dynamics, nonlinear time series analysis, and complex system modeling. His recent works have been on the application of mathematical methods to a diverse range of problems, including data science, biomathematics and the interpretability of the deep learning model.



**Guangming Lu** received the B.S. degree in electrical engineering, the M.S. degree in control theory and control engineering, and the Ph.D. degree in computer science and engineering from the Harbin Institute of Technology, Harbin, China, in 1998, 2000, and 2005, respectively. He is currently a Professor with Harbin Institute of Technology, Shenzhen, China. His current research interests include pattern recognition, image processing, and automated biometric technologies and applications.