2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Cause-based Supervised Contrastive Learning with Adversarial Sample-label for Multimodal Emotion Recognition in Conversation

Yujie Guan, Yu Wang, Weijie Feng and Tingyi Li and Xiao Sun (✉)

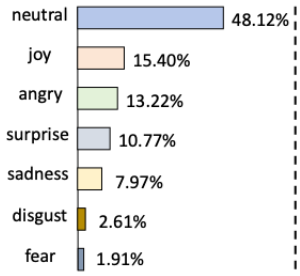
School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China
sunx@hfut.edu.cn

Abstract. Multimodal Emotion Recognition in Conversation (MERC) holds significant importance in Natural Language Processing due to its broad range of applications. However, existing methods still face challenges in addressing the highly imbalanced class problem and extracting robust representations for complex conversational scenarios. To address these challenges, a cause-based supervised contrastive learning framework with adversarial sample-label (CaSCLA) is proposed in this paper. Specifically, we employ a modality balancing technique to fuse the multimodal features, which are then fed into a novel causal-aware network to effectively capture the underlying causal relationships within dialogues. Besides, a supervised contrastive learning with adversarial sample-label method is proposed to alleviate the class imbalance problem by learning label representations and optimizing the similarity between sample features and label embeddings. Furthermore, CaSCLA applies an adversarial samples training strategy, constructing additional positive sample-label pairs to enhance the diversity of the data and increase the robustness of the model. Extensive experiments on the IEMOCAP and MELD benchmark datasets demonstrate that CaSCLA achieves competitive performance.

Keywords: Class Imbalance, Supervised Contrastive Learning, Emotion Recognition in Conversation.

1 Introduction

Multimodal emotion recognition in conversation (MERC) is crucial for dialogue systems, enabling the understanding and generation of empathetic responses [21]. It has widespread applications in areas such as social media analysis, empathetic chatbots, and medical diagnosis. Previous studies primarily focus on context modeling [30] and emotion representation learning [14, 29]. Recently, there has been a growing body of research [10, 25] that leverages supervised contrastive learning (SCL) to learn robust and generalized feature representations better suited for emotion classification.



(a) Emotional ratio

Methods	Meld		
	fear	disgust	WAF1
ConGCN	08.70	10.60	59.40
DialogueRNN	01.23	01.73	55.97
CTNet	10.00	11.20	60.50
MM-DFN	01.02	00.00	57.85
CauAIN	12.90	20.81	64.84
MVTCN	08.80	06.70	58.60

(b) Classification effect

Fig. 1. An illustrative example of imbalanced data distribution on the MELD benchmark dataset. (a) The ratios of the seven emotion labels to all, respectively. (b) The overall performance of some baseline models and their classification performance on the minority class labels "fear" and "disgust".

Although these methods have yielded some progress, they still face certain limitations: **(1)** Existing methods often overlook the severe class imbalance issues in widely used ERC benchmark datasets. This leads to a satisfactory overall performance, but notably poor F1 scores for minority classes. Taking the popular multimodal benchmark dataset MELD [20] shown in **Fig. 1** as an example, the “disgust” and “fear” emotion labels only account for 2.61% and 1.91% of the total labels, respectively, and all baseline models achieve F1 scores below 20.81% for these two emotion labels. Similarly, this problem exists on other multimodal benchmark datasets. **(2)** In the domain of representation learning, label-based contrastive learning techniques generate generalized representations by capturing intra-class similarities while contrasting them with inter-class instances. However, existing methods often fail to fully exploit emotion label information, hindering the discrimination of similar emotions in complex dialogue scenarios. **(3)** When modeling the dependencies of intra-speaker emotional changes and inter-speaker emotional interactions, there is a limitation in capturing deeper and richer emotional dynamics clues due to not considering the actual emotional triggers that elicit the target emotions.

To address the aforementioned issues, we propose an innovative approach named CaSCLA, which balances multimodal data and retrieves causal clues from commonsense knowledge to guide the causal feature extraction process. It is a modified version of Causal Aware Interaction Network (CauAIN) [32]. Additionally, the SCLA module constructs positive and negative sample-label pairs and incorporates multimodal information along with adversarial training, effectively addressing class imbalance issues while enhancing the model's robustness and generalization ability. Another significant contribution of our method is the application of the adversarial samples training (AST) strategy through Soft SCL [6], which generates context-aware perturbations to create hard positives, thereby dispersing the representation space of each class and challenging less robust models. Besides, we conduct extensive experiments on the IEMOCAP and MELD benchmark datasets to confirm the superiority of our method. The main contributions of this work can be summarized as follows:

- We propose a cause-based supervised contrastive learning framework with adversarial sample-label, named CaSCLA, which not only fuses semantic information across modalities, but also utilizes a novel SCLA module to alleviate the class imbalance issue.
- We introduce a causal aware interaction network to accurately recognize complex emotions in conversations from the perspective of emotional causes.
- We devise an AST strategy that enhances the model's contextual robustness by generating context-level worst-case samples as adversarial samples.
- The results of extensive experiments on two datasets demonstrate the effectiveness of CaSCLA, especially on the emotion of the minority class.

2 Related Work

2.1 Emotion Recognition in Conversation

The field of ERC has made significant strides over the past decade, with achievements broadly classified into unimodal and multimodal approaches. Unimodal approaches tend to solve the emotion recognition task using only text modality. DialogueRNN [17] considers the global state, party state, and emotion state of speakers and utilizes three GRUs to model intra- and inter-speaker dependency. DAG-ERC [24] employs directed acyclic graphs to represent the interaction between utterances and speakers. COGMEN [9] is a contextualized graph neural network that integrates both local and global conversational data. CoMPM [11] utilizes a transformer-encoder and a speaker-aware pretrained memory module to capture conversational contexts. Over recent years, MERC integrates multiple sensory channels to boost the model's emotional perception. LR-GCN [22] introduces a latent relationship representation learning mechanism to represent the interactive relationships by learning the latent connections between nodes. GraphMFT [12] utilizes multiple enhanced graph attention networks to capture intra-modal contextual information and inter-modal complementary information. While these methods help capture information in multimodal data, they do not specifically address the challenge of class imbalance.

2.2 Imbalanced Classification

Some studies have observed the issue of class imbalance in MERC. Dave and Khare [3] tackle this issue by using class weights to fine-tune the model training process, which is a commonly adopted methodology. However, selecting appropriate weights relies on domain expertise, and improper choices can degrade performance or increase instability. CBERL [18] combines data augmentation, deep joint variational autoencoders, and multi-task graph neural networks to enhance model classification performance and mitigate the impact of class imbalance. However, it fails to account for the characteristics of multimodal conversational emotion recognition data and overlooks the misclassification of correct samples caused by decision boundary adjustments. This may reduce the model's generalization ability and hinder the recognition of minority emotion categories.

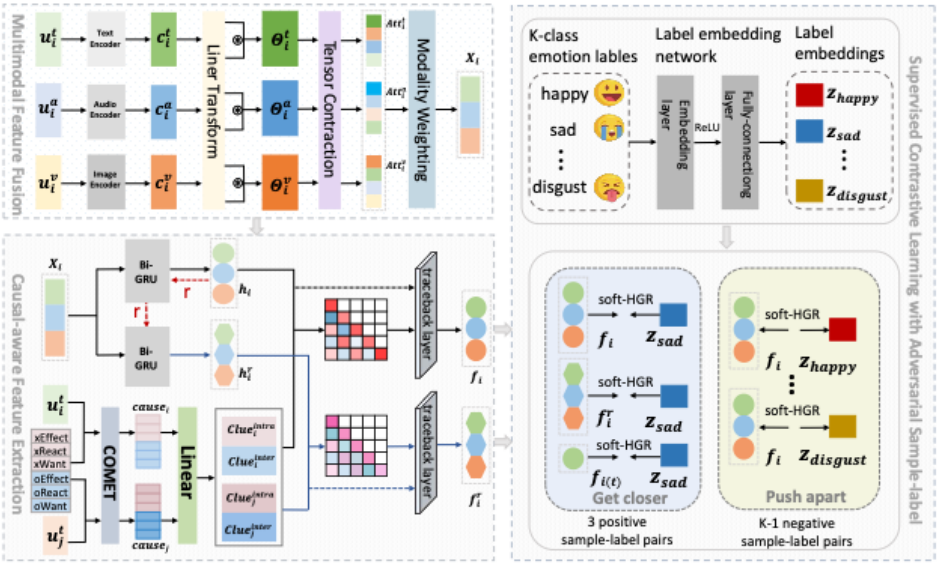


Fig. 2. The overall architecture of CaSCLA.

3 Methodology

Formally, a dialogue contains a sequence of M utterances $\{u_1, u_2, \ldots, u_M\}$, where each utterance u_i consists of multi-sensory data, including textual u_i^t, acoustic u_i^a and visual u_i^v modalities. Predicting the emotion category y_i of each utterance in a dialogue from a predefined set of K classes is the goal of MERC.

Fig. 2 shows the architecture of the CaSCLA, which consists of three modules. In the multimodal feature fusion module, we infer feature-aware attention weights for each modality, which aims to produce a balanced representation. The causal-aware interaction module utilizes commonsense knowledge to capture deep emotional clues and derive causal features. In the SCLA module, we adopt a novel label learning approach and constructing positive and negative sample-label pairs for contrastive learning. We will discuss each component in detail as follows.

3.1 Multimodal Feature Fusion

To retain the unique characteristics of each modality, we create corresponding feature extractors: RoBERTa [16] for text modality and openSMILE toolkit for audio modality. For visual features, we use a pretrained DenseNet for the MELD dataset and a 3D-CNN for the IEMOCAP dataset. Subsequently, we utilize a Transformer [26] network as the encoder to generate a unimodal representation respecting to the modality m as:

$$c_i^m = \phi\left(\theta^{(m)}, u_i^m\right), m \in \{t, a, v\} \tag{1}$$

Where $c_i^m \in R^{M\times d_m}$, the function $\phi(\theta^{(m)})$ is the Transformer network with learnable parameter $\theta(m)$.

Given the issue of modality imbalance in multimodal ERC, we apply tensor contraction [31] to infer feature-aware attention weights for each modality. Specifically, we extend the traditional attention mechanism by replacing the query (Q) and key (K) representations with tensor-ring decomposition-based counterparts. The modality m core tensor $\mathcal{G}_{i,\mathcal{Q}}^m \in R^{d_m\times r_s\times r_w}$ and $\mathcal{G}_{i,\mathcal{K}}^m \in R^{d_m\times r_s\times r_w}$ are computed using a Linear Transform , as expressed below:

$$\begin{cases} \mathcal{G}_{i,\mathcal{Q}}^m = \text{reshape}\left(\left(c_i^m W_{Q_m}^{(1)}\right) \otimes_1 \left(c_i^m W_{Q_m}^{(2)}\right)\right) \\ \mathcal{G}_{i,\mathcal{K}}^m = \text{reshape}\left(\left(c_i^m W_{K_m}^{(1)}\right) \otimes_1 \left(c_i^m W_{K_m}^{\{(2)\}}\right)\right) \end{cases} \tag{2}$$

where $m \in \{t,a,v\}$, $W_{Q_m}^{(1)} \in R^{d_m\times r_s}$, $W_{Q_m}^{(2)} \in R^{d_m\times r_w}$, $W_{K_m}^{(1)} \in R^{d_m\times r_s}$, $W_{K_m}^{(2)} \in R^{d_m\times r_w}$ are the linear transformation matrices, the index $s,w \in \{1,2,3\}$, and $s \neq w$; $\otimes_1$ denotes the mode-1 Khatri-Rao product. To perform multimodal attention in the tensor space, we need to compute the attention coefficient matrix $\Theta_i^{(m)}$ from the tensorized input:

$$\Theta_i^m = \text{softmax}\left(\frac{1}{\sqrt{d_k}}\mathcal{G}_{i,\mathcal{Q}}^m \odot \mathcal{G}_{i,\mathcal{K}}^m\right) \tag{3}$$

where $\odot$ denotes the element-wise product, $\sqrt{d_k}$ is a scaling factor. Then, we use attention mechanism to address the varying impact of each modality on inter- and intra-modality interactions. The attention pooling matrices $\mathbf{A}_i^{(m)} \in R^{r_s\times r_w}$ can be calculated by averaging $\Theta_i^{(m)}$, and the feature-aware attention matrix $Att_i^m \in R^{M\times d_m}$ is computed as follows:

$$Att_i^m = \text{Linear}\left(\Theta_i^{(m)} \times_3^1 \mathbf{A}_i^{(t)} \times_3^1 \mathbf{A}_i^{(a)} \times_3^1 \mathbf{A}_i^{(v)}\right) \tag{4}$$

where $\times_3^1$ is the mode - $(_3^1)$ tensor contraction. The feature-aware balanced representation $x_i^{m(f-adapt)} \in R^{M\times d_m}$ for a given modality m is computed as:

$$x_i^{m(f-adapt)} = Att_i^m c_i^m + \eta c_i^m \tag{5}$$

where $\eta \in [0,1]$ is a parameter to regulate the contribution of the original unimodal feature vector c_i^m. Finally, we incorporate modality-wise L2 normalization to properly weight features to prevent any single modality from dominating the fusion process. The final multimodal balanced representation $\mathbf{x}_i$ of the given utterance is calculated as follows:

$$\mathbf{x}_i = \sum_{m\in\{t,a,v\}} \frac{W^m x_i^{m(f-\text{adapt})}}{\|W^m\|\left\|x_i^{m(f-\text{adapt})}\right\|} \tag{6}$$

where $W^m \in R^{d_m \times |\mathbf{Y}|}$ is the output matrix of the model pertaining to modality m, and **Y** is the set of emotion classes.

3.2 Causal-aware Feature Extraction

Currently, common ERC datasets lack emotion cause labels. Previous works [23] demonstrate that neural networks can anticipate causes and effects of previously unseen events by leveraging ATOMIC's rich inferential knowledge. ATOMIC is a large-scale knowledge graph composed of structured triplets detailing events, relations, and outcomes. In light of this, we ingeniously leverage ATOMIC to obtain causal clues by exploring six types of these relationships. Among these, $xReact$, $xEffect$ and $xWant$ provide intra-cause clues that represent influences generated by the speaker's own utterances, while $oReact$, $oEffect$, and $oWant$ capture the effects on others, offering inter-cause clues. We concatenate utterance text features together with relationships and mask tokens (e.g., $(u_i^t\ [MASK]\ oReact)$), and input them into the generative commonsense transformer model COMET [1]. The hidden state representation from the last encoder layer of COMET is taken as a causal clue. For each u_i, three intra-cause clues are concatenated and passed through a linear layer to obtain a $2d_h$-dimensional $Clue_i^{intra}$. While the inter-cause clues can be similarly represented as $Clue_i^{inter}$.

In the conversation scenario, contextual information can aid in comprehending the emotion of the current utterance. We leverage multimodal features $\mathbf{x}_i$ and the Bi-directional Gated Recurrent Unit (Bi-GRU) to model the sequential dependencies between utterances. The original contextual representation h_i and its adversarial contextual representation h_i^r are computed as in Eq.7. The h_i^r is obtained by introducing minor perturbations within the Bi-GRU, and the details on perturbation generation will be provided in Section 3.4.

$$h_i = \overleftrightarrow{GRU}(\mathbf{x}_i, h_{i-1}), h_i^r = \overleftrightarrow{GRU}(\mathbf{x}_i, h_{i-1}^r) \tag{7}$$

where $h_i \in R^{2\times d_h}$ represents the hidden state vector at time step i and d_h is the dimension of a GRU cell output.

To further enrich context representation with emotion cause, we establish the two-step causal-aware interaction to extract causal features. Initially, we retrieve causal clues and assign corresponding weights to explore the relationship between the emotional reasons of the target utterance and intra- or inter-cause utterances. When determining the association degree of intra-cause clues, we focus on the same speaker's impact and evaluate the association score $s_{i,j}^{intra}$:

$$s_{i,j}^{intra} = \frac{\left[f_q(h_i)\left(f_k(h_j)+f_e\left(Clue_j^{intra}\right)\right)\right]m_{i,j}^{intra}}{\sqrt{d_h}} \tag{8}$$

$$m_{i,j}^{intra} = \begin{cases} 1, j \le i \text{ and } \phi(h_i) = \phi(h_j) \\ 0, \text{ otherwise} \end{cases} \tag{9}$$

where $f_q(\cdot)$, $f_k(\cdot)$, $f_e(\cdot)$ are all linear transformations, ϕ maps the index of the utterance into that of the corresponding speaker, $m_{i,j}^{intra}$ checks whether the retrieved causal clue and the target utterance originate from the same speaker. Similarly, we can compute the association degree of inter-cause clues $s_{i,j}^{inter}$ by focusing on utterances from other speakers. At the causal utterance traceback layer, we perform weighted aggregation of the contextual representation h_i and the emotional clues from the preceding layer, guided by the association scores to emphasize emotion-relevant utterances, yielding the final causal-aware feature representation f_i. Correspondingly, the final adversarial feature representation f_i^r is produced through the causal-aware interaction between the adversarial contextual representation h_i^r and the causal clues.

3.3 Supervised Contrastive Learning

Supervised contrastive learning is a representation learning technique that aims to enhance the robustness of learned representations. However, when dealing with imbalanced datasets, a larger batch size is required to ensure that each minority class sample has at least one positive pair within the batch. The SCLA module proposed in this paper offers a novel solution to this problem.

Similarity Measure. We embed each discrete emotion category into a dense label embedding using a simple yet effective two-layer MLP. Specifically, given a K-class emotion category set $\mathcal{Y} = \{1,2,\dots,K\}$, the label embedding $z_i \in R^d$ for the i-th emotion label is obtained as follows:

$$z_i = \text{Dropout}\left(\text{ReLu}\left(W_g e_i + b_g\right)\right) \tag{10}$$

$$e_i = \max\left(0,\text{EmbeddingLayer}(i)\right) \tag{11}$$

where $e_i \in R^{d_e}$ is the hidden output, $W_g \in R^{d \times d_e}$ is the weight matrix, and $b_g \in R^d$ is the bias parameter. Subsequently, we utilize the label embedding z_i as the ground-truth representation for class i samples. In contrast to traditional methods, our training objectives are designed to maximize the similarity between sample feature and its ground-truth label embedding, while minimizing similarity with label embedding from different classes. This alteration ensures that each sample has at least one positive pair within the batch, regardless of the batch size. In addition, we compute the similarity using soft-HGR [27]. Give a batch of M samples, we obtain the feature representations $F = [f_1, f_2, \cdots, f_M]^{\mathrm{T}} \in R^{M \times d}$, and the sentiment labels $Y = [y_1, y_2, \cdots, y_M]^{\mathrm{T}} \in R^M$, are mapped to obtain the label embeddings $Z_y = \left[z_{y_1}, z_{y_2}, \cdots, z_{y_M}\right]^{\mathrm{T}} \in R^{M \times d}$. We employ the inner product formed from feature covariances as a soft regularizer, thereby replacing the hard whitening constraints in HGR. Then the similarity of a sample-label pair $f_i \in F$ and $z_{y_i} \in Z_y$ can be expressed as:

$$\text{Sim}\left(f_i, z_{y_i}\right) = \frac{1}{M-1} f_i^{\mathrm{T}} z_{y_i} - \frac{1}{2}\sum_{l=1}^{M} cov\left(f_i, f_l\right)\text{cov}\left(z_{y_i}, z_{y_l}\right) \tag{12}$$

Positive Sample-Label Pairs Generation. When y_i is the ground-truth label of f_i, we treat (f_i, z_{y_i}) as a positive sample-label pair. However, for a minority of samples, there is only one positive sample-label pair, which may prevent the model from adequately learning the diversity within categories, thereby restricting its generalization ability.

To tackle this limitation, inspired by adversarial training [19], a regularization method for models to improve robustness, we introduce adversarial perturbations to generate additional positive sample f_i^r. On the one hand, we construct the adversarial sample-label pairs as augmentations of the original sample features, on the other hand, incorporating such challenging samples into the training process facilitates the enhancement of the model's robustness. Additionally, given the crucial role of the textual modality in emotion classification, we also incorporate textual causal-aware features $f_{i(t)}$, obtained through causal interaction from textual features c_i^t, as augmentations of f_i. Importantly, our data augmentation technique does not alter the multimodal fusion mechanism employed in our model. The augmented positive sample-label loss for the i-th sample can be calculated as follows:

$$\mathrm{Aug}^{(i)} = \sum_{\tilde{f} \in F_{\mathrm{aug}}^{(i)}} \exp\left(\mathrm{Sim}\left(\tilde{f}, z_{y_i}\right)\right) \tag{13}$$

$$p_{\mathrm{pos}}^{(i)} = \frac{\exp\left(\mathrm{Sim}\left(f, z_{y_i}\right)\right)}{\sum_{z_j \in \mathbf{Z}} \exp\left(\mathrm{Sim}\left(f_i, z_j\right)\right) + \mathrm{Aug}^{(i)}} \tag{14}$$

$$l_{\mathrm{pos}}^{(i)} = -\sum_{f \in F_{\mathrm{aug}}^{(i)} \cup \{f_i\}} \log\left(p_{\mathrm{pos}}^{(i)}\right) \left(1 - p_{\mathrm{pos}}^{(i)}\right)^{\alpha} \tag{15}$$

where α is a positive focusing parameter that forces the model to focus on hard positive examples, $F_{\mathrm{aug}}^{(i)}$ is the set of augmented sample features for f_i and comprises $\{f_{i(t)}, f_i^r\}$, $Z = [z_1, z_2, \cdots, z_K]^{\mathrm{T}} \in R^{K \times d}$.

Negative Sample-Label Pairs. A pair (f_i, z_{y_i}) is referred to as a negative sample-label pair if y_i is not the ground-truth label of f_i. Unlike existing SCL approaches that do not directly compute the loss from negative pairs, SCLA loss includes a term specifically for negative pairs. If the model incorrectly predicts samples, this term penalizes the model to minimize the similarity between such pairs. The negative sample-label loss for the i-th sample can be calculated as follows:

$$p_{\mathrm{neg}}^{(i)} = \frac{\exp\left(\mathrm{Sim}\left(f_i, z_{y_i}\right)\right)}{\sum_{z_j \in \mathbf{Z}} \exp\left(\mathrm{Sim}\left(f_i, z_j\right)\right)} \tag{16}$$

$$l_{\mathrm{neg}}^{(i)} = -\sum_{z_{y_i} \in Z_{\mathrm{neg}}^{(i)}} \log\left(1 - p_{\mathrm{neg}}^{(i)}\right) {p_{\mathrm{neg}}^{(i)}}^{\beta} \tag{17}$$

where β is a negative focusing parameter that assigns more focus to hard negative samples, $Z_{\text{neg}}^{(i)} = \{z_k \in \boldsymbol{Z} | z_k \neq z_{y_i}\}$ is the set of negative label embeddings for f_i.

In summary, SCLA loss for M samples can be calculated as follows:

$$L_{\text{SCLA}} = \sum_{i=1}^{M} \left(\frac{M}{n_{y_i}}\right)^{\gamma} \left(l_{\text{pos}}^{(i)} + l_{\text{neg}}^{(i)}\right) \tag{18}$$

where n_{y_i} is the count of label y_i within the batch, γ is a parameter to assign higher weights to minority emotions.

3.4 Adversarial Samples Training

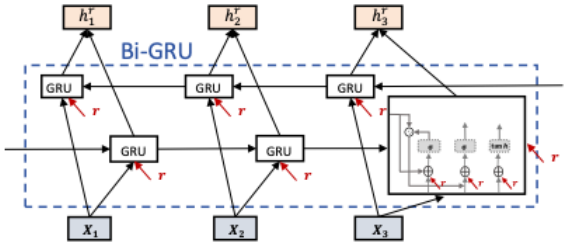


Fig. 3. A Bi-GRU network with contextual adversarial perturbations $\boldsymbol{r}$.

Adversarial training [5] is a technique designed to bolster model robustness to small, approximately worst-case perturbations by incorporating adversarial examples into the training process. In context-dependent scenarios, directly generating adversarial samples can disrupt sample relationships and impair context comprehension. To solve this problem, we design a contextual adversarial sample training (AST) strategy based on Bi-GRU to obtain diverse contextual features and enhance model robustness. Let us denote $\mathbf{x}_i$ as the input, θ as the parameters of the model and ϵ as the norm constraint to limit the size of the adversarial perturbations. When applied to a classifier, adversarial training adds the following term to the cost function $-\log p(y \mid \mathbf{x}_i + \boldsymbol{r}_{adv}; \theta)$, where

$$\boldsymbol{r}_{adv} = \arg\min_{\boldsymbol{r}, |\boldsymbol{r}| \leq \epsilon} \log p\left(y \mid \mathbf{x}_i + \boldsymbol{r}; \hat{\theta}\right) \tag{19}$$

in which r is a perturbation on the input and $\hat{\theta}$ is a constant set to the current parameters of a classifier. At each step of training, we identify the worst case perturbations $\boldsymbol{r}_{adv}$ against the current model $p(y \mid \mathbf{x}_i; \hat{\theta})$ in Eq. (19), and train the model to be robust to such perturbations through minimizing Eq. (19) with respect to θ. However, this value is difficult to calculate accurately, because exact minimization with respect to $\boldsymbol{r}$ is impractical for complex models like neural networks. [5] proposed to approximate it by linearizing $\log p\left(y \mid \mathbf{x}_i; \hat{\theta}\right)$ around $\mathbf{x}_i$. With a linear approximation and a L_2

norm constraint in Eq. (19), the resulting adversarial perturbation shown by Eq. (20) is put on the context-aware hidden layers of the model.

$$\boldsymbol{r}_{adv} = -\epsilon g/\|g\|_2 \text{ where } g = \nabla_{\mathbf{x}_i} \log p\left(y \mid \mathbf{x}_i; \hat{\theta}\right) \tag{20}$$

Here, we take the Bi-GRU network with a sequence input $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M]$ as an example. Adversarial perturbations are put on context-aware hidden layers of the GRU in a multi-channel way, including two gated layers and candidate hidden state in the GRU structure, as shown in **Fig. 3**. After introducing AST, we obtain more diverse adversarial sample features h_i^r. Most importantly, our SCLA can construct more positive sample-label pairs to enable the model to learn more smooth representation spaces from context-dependent inputs, as well as enhance the model's context robustness.

3.5 Overall Training Objective

The overall training objective is a linear combination of the SCLA loss and the cross-entropy (CE) loss, which is defined as follows:

$$L_{Train} = L_{CE} + L_{SCLA} \tag{21}$$

$$L_{CE} = -\sum_{i=1}^{M} p_i[y_i] \tag{22}$$

where p_i is the probability distribution of emotion classes for f_i.

4 Experiments

4.1 Dataset

Table 1. The statistics of two datasets.

Datasets	Dialogues			Utterances		
	train	val	test	train	val	test
IEMOCAP	120		31	5810		1623
MELD	1039	114	280	9989	1109	2610

We evaluate model on two benchmark ERC datasets. The statistics are reported in **Table 1**. IEMOCAP [2] is a multimodal ERC dataset in which each conversation is attended by two speakers. Each utterance is labeled as one of six emotions, namely neutral, happy, sad, anger, frustrated, and excited. MELD [20] is a multi-modal multi-party conversation dataset from the TV series. There are seven emotion labels: neutral, joy, surprise, sadness, anger, disgust, and fear.

Table 2. Overall results against various methods on IEMOCAP.

Models	Happy	Sad	Neutral	Angry	Excited	Frustrated	W-f1
DialogueRNN*	33.18	78.80	59.21	65.28	71.86	58.91	62.75
DialogueCRN‡	57.69	**82.35**	60.63	63.57	75.63	58.66	66.26
CTNet*	51.30	79.90	65.80	67.20	78.70	58.80	67.50
MM − DFN‡	38.62	80.59	67.27	**69.11**	75.29	66.81	67.90
SPCL − CL†	-	-	-	-	-	-	67.74
CauAIN†	60.29	74.14	62.90	63.31	70.61	64.00	66.09
CoG − BART*	-	-	-	-	-	-	66.18
MVTCN*	30.20	74.20	59.00	62.70	72.50	66.60	64.10
DialogueINAB*	-	-	-	-	-	-	67.22
CaSCLA	**65.32**	76.11	**67.71**	62.37	**79.12**	**67.24**	**68.56**

Table 3. Overall results against various methods on MELD.

Models	Neu-tral	Sur-prise	Fear	Sad	Joy	Dis-gust	Anger	W-f1
ConGCN*	77.40	50.30	08.70	28.50	53.10	10.60	46.80	59.40
DialogueRNN*	73.54	49.47	01.23	23.83	50.74	01.73	41.54	55.97
DialogueCRN‡	77.03	47.44	00.00	21.07	52.95	00.00	46.97	58.22
CTNet*	77.40	52.70	10.00	32.50	56.00	11.20	44.60	60.50
MM − DFN‡	75.30	48.42	01.02	26.87	53.31	00.00	45.63	57.85
SPCL − CL†	-	-	-	-	-	-	-	65.45
CauAIN†	**78.91**	**57.95**	12.90	41.65	61.44	20.81	52.20	64.84
CoG − BART*	-	-	-	-	-	-	-	64.81
MVTCN*	76.00	46.90	08.80	04.60	52.30	06.70	47.60	58.60
DialogueINAB*	-	-	-	-	-	-	-	57.78
CaSCLA	76.29	55.37	**27.72**	**44.34**	**62.51**	**31.78**	**54.09**	**66.21**

NOTE: Models annotated with the ∗ symbol indicate results are from their original papers, the † symbol means that the results are obtained through author replication, and the ‡ symbol denotes results from [13].

4.2 Baselines and Implementation Details

We compare our proposed model with the following methods: ConGCN [30], DialogueRNN [17], DialogueCRN [8], CTNet [15], MM-DFN [7], SPCL-CL [25], CauAIN [32], CoG-BART [14], MVTCN [28] and DialogueINAB [4]. We choose F1 score per class and weighted-F1 score as the evaluating metric. All of our results are averaged on 5 runs.

Hyperparameter Settings: For unimodal feature extraction, models are trained using Adam optimizer with a batch size of 32 and the dimensionality of the unimodal feature vectors d_m is 1024 on both datasets. For all representations in the following parts of CaSCLA, d_h is set to 300. The training of CaSCLA is also performed using the Adam optimizer, with a learning rate of 1e-4. The positive and negative focusing parameters α and β are set to 2.0 and 0.5 respectively, the sample-weight parameter γ is chosen among {0.5, 1.0, 1.5}.

5 Results and Analysis

5.1 Overall Results

The overall results are reported on two datasets, IEMOCAP and MELD, in **Table 2** and **Table 3** respectively. CaSCLA achieves the best overall recognition performance across the two datasets compared to other methods. Specifically, on the IEMOCAP dataset, CaSCLA achieves a weighted F1 score of 68.56%, marking a 2.47% improvement over the baseline model CauAIN and an average gain of 2.56% compared to other state-of-the-art models. On the MELD dataset, our model attains a weighted F1 score of 66.21%, outperforming the baseline CauAIN by 1.37% and achieving an average improvement of 6.44% over other state-of-the-art methods.

We also report fine-grained results on two datasets. It is noteworthy that CaSCLA achieves better results for most emotion categories (5 out of 7 classes) on the highly imbalanced MELD dataset. Particularly for the minority categories of "fear" and "disgust", its F1 scores surpass the second-best model CauAIN, by 14.82% and 10.97% respectively, while showing average improvements of 20.61% and 23.27% over other models. Additionally, on the IEMOCAP dataset, for the happy category, which is the least represented among all categories and accounts for only 9% of the total categories, CaSCLA achieves the highest F1 score of 65.32%, representing an average improvement of 20.11% over other models, while also demonstrating strong performance across the remaining categories.

These results indicate that CaSCLA not only achieves the best overall performance but also makes significant improvements in minority classes, effectively addressing the class imbalance. Taking all aspects into consideration, our model demonstrates the most outstanding performance in emotion recognition.

5.2 Ablation Study

We conduct ablation studies, with results presented in **Table 4**.

Impact of SCLA. To study the contribution of SCLA, we remove the proposed SCLA module. The experimental results show that the performance of $\text{CaSCLA}_{\text{w/o SCLA}}$ drops significantly on both datasets. These demonstrate that the SCLA module effectively optimizes the model's performance in supervised contrastive learning tasks by constructing positive and negative sample label pairs. It not only enhances the model's

Table 4. Ablation results of CaSCLA.

Models	Performance(weighted-F1)	
	IEMOCA	MELD
CaSCLA	68.56	66.21
w/o SCLA	66.89(↓1.67)	64.97(↓1.24)
w/o AST	67.23(↓1.33)	65.35(↓0.86)
w/o causal-aware	67.31(↓1.25)	65.32(↓0.89)

ability to distinguish between different classes but also improves its handling of challenging samples.

Impact of AST. We implement $\text{CaSCLA}_{\text{w/o AST}}$, excluding adversarial training for generating additional sample features. Its performance significantly deteriorates on both datasets, with the most pronounced decline observed on IEMOCAP, likely due to the presence of semantically similar emotion categories, such as "happy" and "excited". This confirms the superiority of AST in enhancing the model's ability to identify similar emotions in complex dialogue scenarios.

Impact of causal-aware. When the causal-aware interaction module is removed (i.e., w/o causal-aware) and the contextual feature representation h_i is directly used as the sample feature, the performance significantly declines on both datasets. This result validates the effectiveness of our approach in detecting emotional causes and improving emotion recognition performance.

5.3 Batch Sizes Stability of CaSCLA

Table 5. Performance comparison on MELD dataset.

Models	MELD (weighted-F1)			
	BS=4	BS=8	BS=16	BS=32
SupCon	57.20	61.43	62.57	64.16
SPCL-CL	61.34	62.92	63.54	65.45
CaSCLA(Ours)	**65.74**	**66.21**	**65.94**	**65.85**

To investigate the stability of CaSCLA across different batch sizes, we conduct experiments on the imbalanced MELD dataset. As shown in **Table 5**, the results show that CaSCLA consistently outperforms methods using these two SCL loss functions across all batch size settings, maintaining superior performance even with small batch sizes. This demonstrates the model's stability across different batch sizes and its effectiveness in addressing large batch sizes limitations.

6 Conclusion

In this paper, to address the limitations of the class imbalance and insufficient model robustness in MERC, we propose an innovative framework named CaSCLA. Specifically, we achieve modality balance and integrate a causal-aware interaction network to extract causal features associated with emotional reasons. Moreover, the SCLA module is designed to map discrete emotion labels into a dense embedding space and construct positive and negative sample-label pairs, thereby alleviating the issue of class imbalance and reducing the necessity of large batch sizes. Additionally, we introduce a contextual AST strategy, which generates additional positive sample features to further enhance the model's robustness. Experimental results demonstrate the effectiveness of CaSCLA.

Acknowledgments. This work was supported by Special Project of the National Natural Science Foundation of China(62441614), Anhui Province Key R&D Program (202304a05020068) and General Programmer of the National Natural Science Foundation of China (62376084)

References

1. Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., Choi, Y.: Comet: Commonsense transformers for automatic knowledge graph construction. arXiv preprint arXiv:1906.05317 (2019)
2. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation 42, 335–359 (2008)
3. Dave, C., Khare, M.: Emotion detection in conversation using class weights. In: 2021 8th International Conference on Soft Computing & Machine Intelligence (IS-CMI). pp. 231–236. IEEE (2021)
4. Ding, J., Chen, X., Lu, P., Yang, Z., Li, X., Du, Y.: Dialogueinab: an interaction neural network based on attitudes and behaviors of interlocutors for dialogue emotion recognition. The Journal of Supercomputing 79(18), 20481–20514 (2023)
5. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
6. Gunel, B., Du, J., Conneau, A., Stoyanov, V.: Supervised contrastive learning for pre-trained language model fine-tuning. arXiv preprint arXiv:2011.01403 (2020)
7. Hu, D., Hou, X., Wei, L., Jiang, L., Mo, Y.: Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp.7037–7041. IEEE (2022)
8. Hu, D., Wei, L., Huai, X.: Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. arXiv preprint arXiv:2106.01978 (2021)
9. Joshi, A., Bhat, A., Jain, A., Singh, A., Modi, A.: Cogmen: Contextualized gnn based multimodal emotion recognition. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4148–4164 (2022)
10. Kim, T., Yoo, K.M., Lee, S.g.: Self-guided contrastive learning for bert sentence representations. arXiv preprint arXiv:2106.07345 (2021)

11. Lee, J., Lee, W.: Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. arXiv preprint arXiv:2108.11626 (2021)
12. Li, J., Wang, X., Lv, G., Zeng, Z.: Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation. Neurocomputing 550, 126427 (2023)
13. Li, Q., Huang, P., Xu, Y., Chen, J., Deng, Y., Yin, S.: Generating and encouraging: An effective framework for solving class imbalance in multimodal emotion recognition conversation. Engineering Applications of Artificial Intelligence 133, 108523 (2024)
14. Li, S., Yan, H., Qiu, X.: Contrast and generation make bart a good dialogue emotion recognizer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 11002–11010 (2022)
15. Lian, Z., Liu, B., Tao, J.: Ctnet: Conversational transformer network for emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29, 985–1000 (2021)
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
17. Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., Cambria, E.: Dialoguernn: An attentive rnn for emotion detection in conversations. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 6818–6825 (2019)
18. Meng, T., Shou, Y., Ai, W., Yin, N., Li, K.: Deep imbalanced learning for multimodal emotion recognition in conversations. IEEE Transactions on Artificial Intelligence (2024)
19. Miyato, T., Dai, A.M., Goodfellow, I.: Adversarial training methods for semisupervised text classification. arXiv preprint arXiv:1605.07725 (2016)
20. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508 (2018)
21. Raamkumar, A.S., Yang, Y.: Empathetic conversational systems: A review of current advances, gaps, and opportunities. IEEE Transactions on Affective Computing 14(4), 2722–2739 (2022)
22. Ren, M., Huang, X., Li, W., Song, D., Nie, W.: Lr-gcn: Latent relation-aware graph convolutional network for conversational emotion recognition. IEEE Transactions on Multimedia 24, 4422–4432 (2021)
23. Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N.A., Choi, Y.: Atomic: An atlas of machine commonsense for if-then reasoning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 3027–3035 (2019)
24. Shen, W., Wu, S., Yang, Y., Quan, X.: Directed acyclic graph network for conversational emotion recognition. arXiv preprint arXiv:2105.12907 (2021)
25. Song, X., Huang, L., Xue, H., Hu, S.: Supervised prototypical contrastive learning for emotion recognition in conversation. arXiv preprint arXiv:2210.08713 (2022)
26. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
27. Wang, L., Wu, J., Huang, S.L., Zheng, L., Xu, X., Zhang, L., Huang, J.: An efficient approach to informative feature extraction from multimodal data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 5281–5288 (2019)
28. Wen, J., Jiang, D., Tu, G., Liu, C., Cambria, E.: Dynamic interactive Multiview memory network for emotion recognition in conversation. Information Fusion 91, 123–133 (2023)
29. Yang, L., Shen, Y., Mao, Y., Cai, L.: Hybrid curriculum learning for emotion recognition in conversation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 11595–11603 (2022)

30. Zhang, D., Wu, L., Sun, C., Li, S., Zhu, Q., Zhou, G.: Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In: IJCAI. pp. 5415–5421. Macao (2019)
31. Zhao, Q., Zhou, G., Xie, S., Zhang, L., Cichocki, A.: Tensor ring decomposition. arXiv preprint arXiv:1606.05535 (2016)
32. Zhao, W., Zhao, Y., Lu, X.: Cauain: Causal aware interaction network for emotion recognition in conversations. In: IJCAI. pp. 4524–4530 (2022)