Deep Imbalanced Learning for Multimodal Emotion Recognition in Conversations

Tao Meng ⓘ, Yuntao Shou ⓘ, Wei Ai ⓘ, Nan Yin ⓘ, and Keqin Li ⓘ, *Fellow, IEEE*

Abstract—The main task of multimodal emotion recognition in conversations (MERC) is to identify the emotions in modalities, e.g., text, audio, image, and video, which is a significant development direction for realizing machine intelligence. However, many data in MERC naturally exhibit an imbalanced distribution of emotion categories, and researchers ignore the negative impact of imbalanced data on emotion recognition. To tackle this problem, we systematically analyze it from three aspects: data augmentation, loss sensitivity, and sampling strategy, and propose the class boundary enhanced representation learning (CBERL) model. Concretely, we first design a multimodal generative adversarial network to address the imbalanced distribution of emotion categories in raw data. Second, a deep joint variational autoencoder is proposed to fuse complementary semantic information across modalities and obtain discriminative feature representations. Finally, we implement a multitask graph neural network with mask reconstruction and classification optimization to solve the problem of overfitting and underfitting in class boundary learning and achieve cross-modal emotion recognition. We have conducted extensive experiments on the interactive emotional dyadic motion capture (IEMOCAP) and multimodal emotion lines dataset (MELD) benchmark datasets, and the results show that CBERL has achieved a certain performance improvement in the effectiveness of emotion recognition. Especially on the minority class "fear" and "disgust" emotion labels, our model improves the accuracy and F1 value by 10% to 20%. Our code is publicly available at https://github.com/yuntaoshou/CBERL.

Impact Statement—MERC plays an important role in human–computer interaction. However, existing methods ignore the data imbalance problem of multimodal datasets. In this article, we propose a class boundary enhanced representation learning (CBERL) model. Since existing multimodal emotion recognition datasets exhibit long-tail distributions on different emotion categories, the proposed method can greatly alleviate the problem of data imbalance and ensure the accuracy of emotion recognition.

In particular, our method can also achieve cross-modal semantic information fusion. Experimental results show that our method outperforms the state of the art (SOTA) methods.

Index Terms—Data augmentation, data imbalance, feature fusion, graph neural network, multimodal emotion recognition in conversations (MERC).

I. INTRODUCTION

WITH the continuous development of hardware resources and social media in recent decades, people have widely used multimodalities, e.g., text, audio, image, and video to express their emotions or thoughts. The task of multimodal emotion recognition in conversations (MERC) is to understand emotions in diverse modalities. It can be widely used in fields such as healthcare, conversation generation, and intelligent recommender systems, which has drawn increasing research attention [1]. For example, in the field of intelligent recommendation, a machine can recommend things that may be most interesting to a consumer based on his changing mood. At the same time, the existence of large multimodal corpus datasets for instant chat software such as Weibo, Meta, and Twitter can provide a data basis for MERC based on deep learning [2]. However, these corpora naturally have a high-class imbalance problem, i.e., most classes only contain a scarce number of samples, while a large number of samples belong to only a few classes.

The current mainstream MERC task mainly uses a graph neural network (GNN) for information fusion to enhance the effectiveness of emotion prediction [3], [4], but they ignore the data imbalance problem. However, in the field of MERC, data imbalance is a widespread problem, which will hinder the model from learning the distribution law of the data and result in the model failing to discern emotion class boundaries. Taking the popular multimodal benchmark dataset multimodal emotion lines dataset (MELD) shown in Fig. 1(a) as an example, the "fear" and "disgust" emotion labels only account for 1.91% and 2.61% of the total labels, respectively, and all baseline models of F1 values are less than 11.2% on the "fear" and "disgust" emotion labels in Fig. 1(b). These emotion classification results cannot meet practical needs. Similarly, this problem also exists on other multimodal benchmark datasets. Therefore, it is necessary to take the data imbalance problem as the starting point of the MERC model design.

To alleviate the data imbalance problem in deep learning, there are mainly three different research directions to

	Neutral	48.12%
	Joy	15.40%
	Angry	13.22%
	Surprise	10.77%
	Sadness	7.97%
	Disgust	2.61%
	Fear	1.91%

(a)

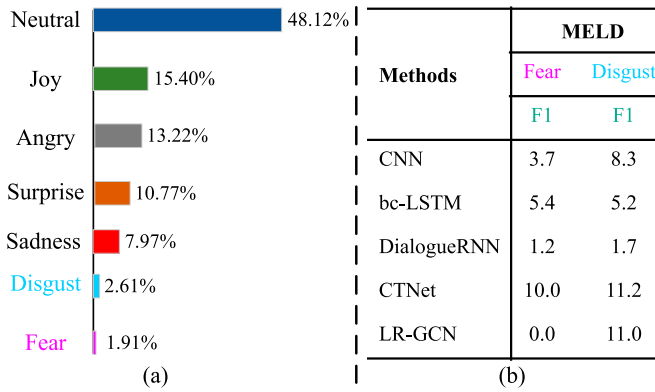| Methods | MELD | |
	Fear	Disgust
	F1	F1
CNN	3.7	8.3
bc-LSTM	5.4	5.2
DialogueRNN	1.2	1.7
CTNet	10.0	11.2
LR-GCN	0.0	11.0

(b)

Fig. 1. Illustrative example of imbalanced data distribution on the MELD benchmark dataset. (a) Ratios of the seven emotion labels to all, respectively. (b) Classification performance of some baseline models on the minority class labels "fear" and "disgust".

optimizing the discriminative degree of class boundaries: data-augmentation level [5], [6], sampling-strategy level [7], [8], and loss-sensitive level [9], [10]. Although these methods have achieved relatively good results in their respective fields, there is still a lack of systematic consideration for the data imbalance problem in MERC.

The research based on data augmentation aims to increase the number of minority class samples to improve the clarity of the model learned class boundaries. For example, Su and Lee [11] proposed the corpus-aware emotional CycleGAN (CAEmoCy-GAN) method, which improves the distribution of generated data through a corpus-aware attention mechanism, so that the model can learn better class boundaries. DA-GAN [12] constructs an image conditional generative adversarial network that uses data from the source domain to learn the data distribution and performs data augmentation by generating intraclass data distribution. Zhu et al. [13] used CycleGAN to generate new image data by complementing and refining the data manifold and finding better margins between adjacent categories. Latif et al. [14] used a hybrid data augmentation scheme to enhance feature learning of GAN and generate speech data. VAE-D2GAN [15] used a variational autoencoder (VAE) architecture to learn the spatial distribution of actual data through latent vectors and introduced it into the dual discriminator GAN to improve the diversity of generated electroencephalography (EEG) data. Although the above methods can achieve relatively good recognition results, they are all dedicated to the generation of unimodal data (i.e., image, EEG, or speech). However, collaboratively generating multimodal data remains a challenging task.

The research based on a sampling-strategy mainly focuses on balancing the ratio of minority-class samples to majority class samples by the sampling frequency. For example, Hamilton et al. [16] utilized GraphSAGE to randomly sample neighbor nodes and exploit their information to generate new node embedding representations. However, such sampling mechanisms may suffer from overfitting or underfitting.

The goal of research based on loss-sensitive function is to make the model learn the distribution of minority class samples by assigning a higher weight to the few samples in the loss

function. For example, Li et al. [10] proposed the gradient harmonization mechanism (GHM), which utilized the gradient density function to balance the weights of the model for easy and hard-to-distinguish samples. However, such methods are susceptible to interference from majority classes or noisy data samples [17].

Overall, the above-mentioned methods have two limitations: 1) Most existing methods perform data augmentation through oversampling or undersampling, and they cannot generate new samples. 2) Although existing GAN-based data augmentation methods can generate new samples, they cannot collaboratively generate multimodal data well due to the heterogeneity problem among multimodal data, resulting in the generated data possibly being noisy data.

In view of the above problems, how to systematically eliminate the negative impact of data imbalance in MERC from three aspects: data-augmentation, sampling-strategy, and loss-sensitive is still a challenging task. Therefore, we propose the class boundary enhanced representation learning (CBERL) model to solve the data imbalance problem in MERC from these three aspects.

The proposed model, CBERL, will first use the data augmentation method of generative adversarial networks (GAN) to generate new samples, thereby providing a data basis for the subsequent model to learn discriminative class boundaries. In the proposed GAN, we design a source-to-target generator and a target-to-source generator. The source-to-target generator is mainly used to learn the original distribution of the data and generate new samples that conform to the distribution of the original data. The target-to-source generator is mainly used to improve the quality of the generated data further. We also introduce identity loss to ensure the consistency of the distribution between the generated data and the original data. Due to the heterogeneity of multimodal data, we introduce emotion classification loss and discrimination loss to guide the generator to learn complementary semantic information and consistent semantic information in multimodal data, and collaboratively generate multimodal data. It is worth noting that the proposed GAN and the subsequent four modules are trained separately. In other words, we perform data augmentation first, and then use the subsequent four modules for emotion classification after completing the data augmentation.

After data augmentation, we input the original and newly generated data into a deep joint variational autoencoder (DJVAE) with KL divergence for cross-modal feature fusion to capture complementary semantic information between different modalities and achieve effective feature dimensionality reduction. Then, we input the fused low-dimensional feature vectors obtained by DJVAE into Bi-LSTM to obtain a feature representation with richer contextual semantic information.

Next, we feed the contextual feature vectors obtained by Bi-LSTM into our proposed multitask graph neural network (MGNN). Specifically, for the first subtask, to overcome the over-fitting or underfitting problems of the random sampling strategy in GNN to the minority class samples, MGNN first randomly performs a mask operation on some nodes in the network during the process of aggregating the information of

surrounding neighbor nodes. Then, the remaining unmasked neighbor nodes are input into the graph convolutional network layer and the multilayer perceptron (MLP), and the predicted values of all neighbor nodes are obtained. Finally, the loss between the true and predicted values is computed to optimize the distribution representation of the class boundaries. For the second subtask, we input the feature vectors extracted by GNN into an emotion classification model composed of multiple weak classifiers and add a conditioning factor to the loss function to enhance the model's propensity to learn the minority class samples. The underlying parameters are shared between the two subtasks of GNN, which helps improve the model's generalization ability, thereby enhancing the performance of emotion recognition.

A. Our Contributions

Therefore, MERC should not only consider the feature fusion problem of text, audio, video, and image modalities, but also generate a new architecture to solve the data imbalance problem. Inspired by the above problems, we propose a novel CBERL model to obtain better emotion class boundaries. The main contributions of this article are as follows.

1) A novel deep imbalanced learning architecture, named CBERL, is presented. CBERL can not only fuse semantic information across modalities, but also learn class boundaries for imbalanced data more accurately.
2) A new generative adversarial network is proposed to generate multimodal samples to provide a data basis for subsequent models to learn class boundaries. The distribution differences between generated and original data and labels are reduced by adding identity loss and classification loss, respectively.
3) We propose a multimodal feature fusion method, i.e., DJVAE. DJVAE estimates the potential distribution of data by introducing KL divergence, so it can learn complementary semantic information between multimodal features and get a more discriminative feature distribution.
4) An MGNN model based on mask reconstruction and classification optimization is proposed to solve the overfitting and under-fitting problems of the random sampling strategy in GNN to the minority class samples.
5) Finally, extensive experiments were conducted on the interactive emotional dyadic motion capture (IEMOCAP) and MELD benchmark datasets. Moreover, compared with the baseline model, CBERL has a better emotion classification effect, especially on the minority class emotion.

II. RELATED WORK

A. Multimodal Emotion Recognition in Conversations

MERC has been widely used in various fields in real life, especially in intelligent dialogue recommendation, and has a high application value. The current mainstream methods mainly focus on three research directions: context-based emotion recognition [18], speaker-based emotion recognition [19], and speaker-distinguishing emotion recognition [20].

In context-based emotion recognition research, Nguyen et al. [21] adopted a deep neural network consisting of a dual-stream autoencoder and a long short-term memory neural network (LSTM), which was able to perform emotion recognition by effectively integrating the conversational context. Qin et al. [22] achieved deep cointeractive relation network (DCR-Net), which interacted with dialogue behaviors and emotion changes by combining BERT's bidirectional encoded representation [23]. Recently, Transformer-based multimodal emotion recognition methods have been proven to be a better technique for modeling long-term contextual correlations [24]. For example, Ma et al. [25] proposed a transformer-based model with self-distillation to capture the intramodal and intermodal information interaction. Lian et al. [19] proposed CT-Net with single Transformer and cross Transformer to extract temporal information and cross-modal semantic information in the discourse.

In speaker-based emotion recognition research, Xing et al. [26] conducted the adapted dynamic memory network (A-DMN), which used a global recurrent neural network (RNN) to model the influence between speakers. However, A-DMN had poor memorization ability on overly long text sequences. Hazarika et al. [27] created the conversational memory network (CMN), which creatively introduced an attention mechanism to obtain the importance of historical context to the current utterance, thereby simulating the dependencies between speakers. However, this method cannot model multidialogue relationships. Ghosal et al. [28] proposed DialogueGCN, which exploited the properties of graph convolutional neural networks (GCN) to construct a dynamic graph model that simulated interactions between speakers by using speakers as nodes of the graph and dependencies between speakers as edges. However, GCNs are prone to over-smoothing, which will cause the model to fail to extract deeper semantic information.

In emotion recognition based on distinguishing speakers, although CMN, ICON, DialogueGCN, and other models modeled the dependencies between different speakers, they did not distinguish who the speaker of the discourse was in the final emotion recognition process. To overcome this problem, Majumder et al. [1] introduced DialogueRNN. The model simultaneously considered the speaker information, the utterance context and the emotional information of multimodal features, and adopted three gated recurrent units (GRU), namely party GRU, global GRU, and emotion GRU, to capture the speaker state, global state of context, and affective state. For the utterance at the current time t, the global state of the context is updated by the context global state at the previous time $t-1$, the context representation at the current time t, and the current speaker's state at the previous time $t-1$. The speaker state was updated by the state of the current speaker at the previous time $t-1$, the representation of the current context, and the global state of the context at the previous time. The affective state was updated by the speaker's current state at time t and the affective state at the previous time $t-1$. Finally, emotion classification is performed with the obtained emotion state.

B. Data Augmentation

The scarcity of datasets has always been an inevitable problem in deep learning and machine learning [29], making it difficult for deep neural network models to learn unbiased representations of real data, resulting in serious overfitting problems. Although regularization methods can alleviate the problem of model overfitting [30], this is not the most essential solution to the problem. Even the simplest machine learning model can achieve very good results when the dataset is large enough. Therefore, we will mainly consider data augmentation methods to improve the model's generalization ability.

Wang et al. [31] adopted deep generative models (DGM), which differed from traditional data enhancement methods by adding Gaussian noise to the original data. DGM designed GAN and VAE conditioned on different input vectors. DGM significantly outperformed traditional audio data enhancement methods. Kang et al. [32] created independent component analysis-evolution (ICA- evolution), which selectively generated data matching the overall data distribution using a fitness function. ICA evolution inherited the idea of a genetic algorithm to enhance the data by crossover and mutation operations. However, this method tended to change the original distribution of the data. Su and Lee [11] proposed Corpus-aware emotional CycleGAN (CAEmoCyGAN), which employed an unsupervised cross-corpus generative model to generate target data with rich semantic a corpus-aware attention mechanism to aggregate important source data information.

C. Feature Fusion Methods

Multimodal features have an important impact on emotion recognition. Feature fusion, as the primary method for information enhancement of multimodal features, has attracted much attention from researchers [33]. Many studies have focused on capturing the differences between modalities to complement multimodal features, and many multimodal feature fusion methods have been successfully employed. Compared with decision-level fusion methods, feature fusion methods can fully use the advantages of multimodal features. Therefore, we will mainly summarize the research related to multimodal feature fusion.

Liu et al. [34] introduced LFM, which used a low-rank tensor method to achieve dimensionality reduction of multimodal features and improve the fusion efficiency. LFM achieved high performance on several different tasks. Zadeh et al. [35] used TFN. The method can learn the semantic information within and between modalities end-to-end. For semantic information extraction between modalities, TFN adopted the method of tensor fusion, which can simulate the interaction process between three modalities of text, audio, and video. TFN can effectively fuse multimodal features. Zhou et al. [36] provided multiscale feature fusion and enhancement (MFFENet), which introduced a spatial attention mechanism to fuse multiscale features with global semantic information. MFFENet could assign higher attention weights to important feature vectors to obtain distinguishable features. Zadeh and Pu [3] proposed DFG, which introduced a dynamic fusion graph model, which can achieve dynamic fusion of multimodal feature vectors, so that various modalities can play complementary roles.

D. Solutions for Imbalanced Data

Despite many vital advances in deep learning in recent decades, imbalanced data is still one of the challenges hindering the development of deep learning models [37]. Therefore, researchers need to design a method to alleviate the problem of sample imbalance. In the existing research, there are three main solutions based on sample sampling, loss function, and model levels.

In a study based on the sample sampling level, Chawla et al. [38] utilized the synthetic minority over-sampling technique (SMOTE) method. It increased the amount of data for a few samples by selecting k neighbors of each minority sample close to each other, then synthesizing each neighboring sample with the original sample manually into a new sample through an equation. However, this algorithm suffered from the marginalization of a small number of samples. Based on the SMOTE algorithm, Han et al. [39] proposed the Borderhne-SMOTE algorithm to increase the data volume of minority class samples by interpolating them in appropriate regions. This method solves the problem of marginalization of thesample distribution. DeepSMOTE can solve the problem of sample imbalance very well. At the level of loss-based function, Lin et al. [9] proposed focal loss, which added a parameter γ to weigh the loss, to balance the contribution of the simple classification samples and the complex classification samples to the loss. Li et al. [10] performed the gradient harmonizing mechanism (GHM), which suppressed the classification results of both simple and difficult classification samples by means of a gradient density function. In model-level-based research, Wang et al. [40] proposed deep-ensemble-level-based Interpretable Takagi-Sugeno-Kang Fuzzy Classifier (DE-TSK-FC), which divided the problem area layer by layer using several successive zero-order TSK fuzzy classifiers and then used K-nearest neighbor (KNN) for classification.

III. METHODOLOGY

A. Design of the CBERL Structure

In this section, we detail the design of the CBERL structure. Fig. 2 visually shows the architecture of the CBERL model proposed in this article. As shown, our model includes five key stages:

1) *Data augmentation:* Unlike previous works that use oversampling or undersampling methods to alleviate the data imbalance problem, we use GAN methods for data augmentation. In addition, unlike traditional GAN methods, we introduce an identity loss to ensure the consistency of the distribution between the generated data and the original data, and emotion classification loss and discrimination loss to guide the generator to capture the complementary and consistent semantic information in multimodal features and collaboratively generate new multimodal data. In particular, we train a generator and a discriminator separately during the data augmentation stage. After completing data augmentation, the model can achieve better emotion classification results. Specifically, We input the newly generated samples together with the
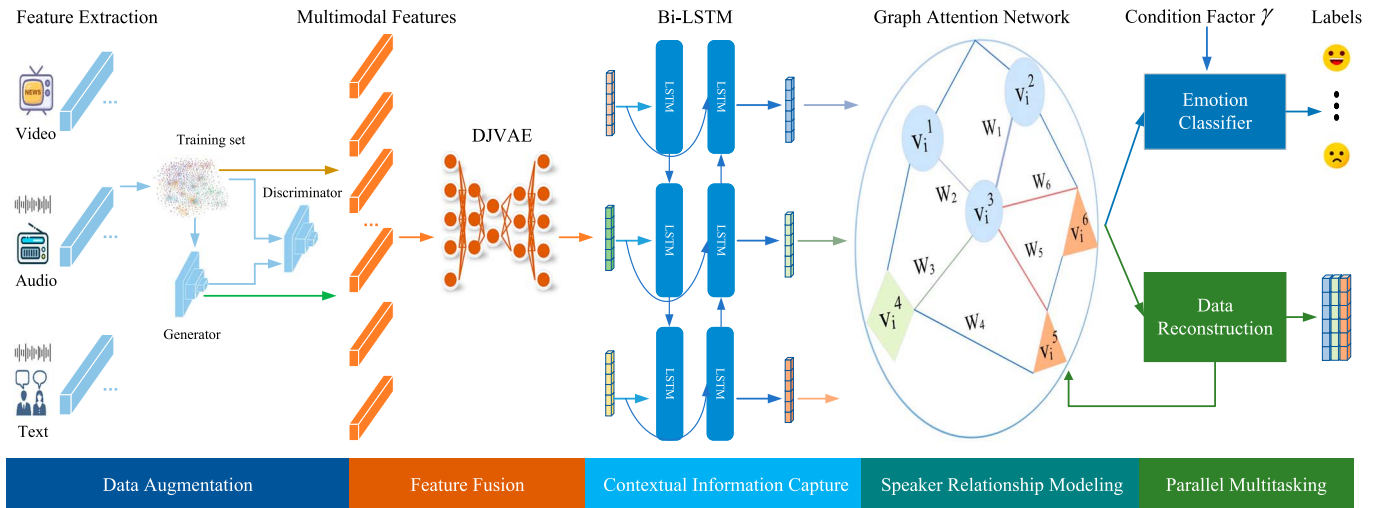
Fig. 2. Framework of the CBERL model consists of a data augmentation layer, an intramodal and intermodal feature fusion layer, a sequential context semantic information extraction layer, a speaker relationship modeling layer, and a emotion classifier.

original samples into the subsequent CBERL model to achieve data balance. It is worth noting that the training of GAN and the subsequent four stages are separate;

2) *Intermodal feature fusion:* After completing data augmentation, we propose a multimodal feature fusion method based on a DJVAE. Different from simple VAE, which only performs point-to-point mapping of raw data, inspired by the idea of joint probability distribution, we introduce KL divergence to estimate the underlying distribution law of raw data, so as to capture the characteristics of differences between modalities, and obtain a more discriminative representation of class boundaries;

3) *Intramodal context feature extraction:* After completing the intermodal feature fusion, we use Bi-LSTM to extract contextual semantic information for the fused text, video and audio features;

4) *Graph interaction:* Increasing evidence shows aggregating the information of all neighboring nodes in the graph will prevent the model from learning the unbiased representation of data from a few classes of nodes. Specifically, we use the contextual semantic features extracted by Bi-LSTM to construct a speaker relationship graph and mask some of the nodes in the graph. Then we utilize GCN to aggregate the information of unmasked nodes. Finally, the semantic information obtained from the aggregation will then perform the data reconstruction and emotion prediction tasks separately to improve the fifitting ability of GCN to the minority class nodes;

5) *Emotion classification:* Finally, we use the feature vectors extracted by GCN to complete emotion classification. Unlike existing MERC methods that utilize fully-connected layers in emotion classification to obtain the final emotion category, we propose a classification optimization algorithm to make the model focus on hard-to-classify samples.

In particular, the data augmentation stage is trained separately. For the four stages of intermodal feature fusion, intramodal contextual feature extraction, graph interaction, and emotion classification, we regard them as a whole for training. Specifically, we first use the video, audio, and text data in the training data set to train a generator and discriminator separately. After the training of the generator and the discriminator is completed, we input the multimodal data generated by the generator that conforms to the original data distribution and the original training data to the feature fusion layer to achieve intramodal and intermodal information fusion. After obtaining the fused multimodal semantic information, we input it into the Bi-LSTM layer to extract the contextual semantic information in the multimodal information. We then use contextual semantic features to construct a speaker relationship graph and use graph convolution operations to aggregate conversational relationship information between speakers. Finally, we utilize the feature vectors extracted after the graph convolution operation to complete the data reconstruction and emotion classification tasks respectively. When performing graph convolution operations, we only use unmasked surrounding neighbor nodes for information aggregation. In particular, in the emotion classification stage, we use cross-entropy loss with a conditioning factor γ for emotion classification.

1) Data Augmentation: To solve the problem of class distribution imbalance in the dataset in MERC, we first build a multisource generator and discriminator for the application scenario of multimodal emotion recognition in conversations. The models then learn the underlying distributions of the multimodal data as they play against each other. Finally, we increase the amount of data required by the model by sampling the data in the learned latent space.

The overview of the data augmentation method is presented in Fig. 3, which consists of a generator G and a discriminator D. In this article, we consider a bidirectional mapping function between source and target data, and use two generators, including a synthetic sample for going from source to target data $(G_{S \to T})$ and a synthetic sample for going from target to source data $(G_{T \to S})$.
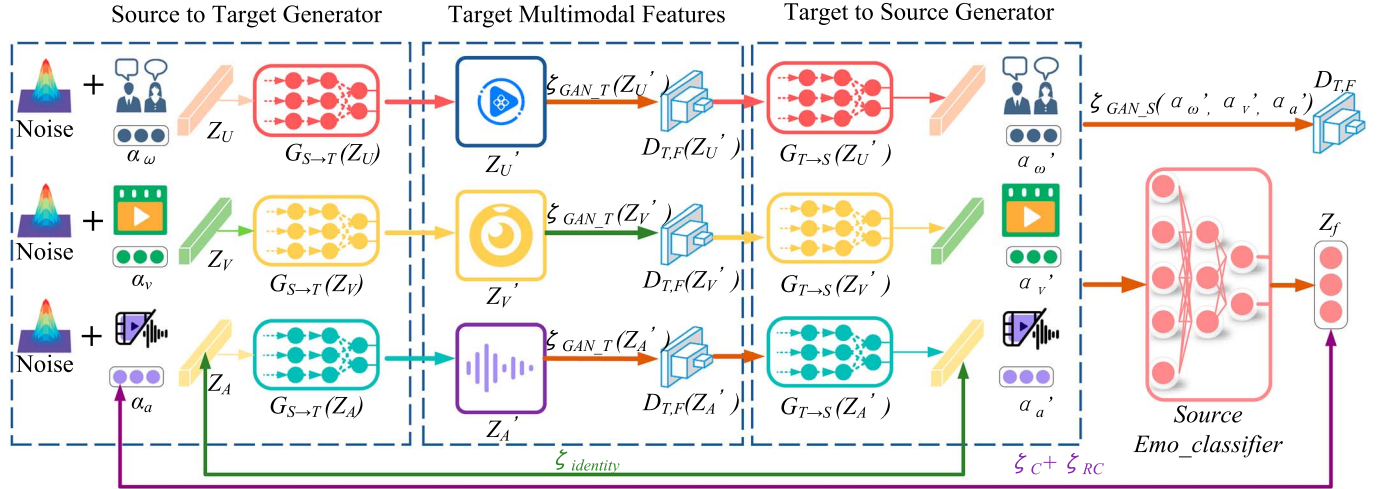
Fig. 3. GAN network consists of a generator, discriminator and emotion classifier. We used two generators for source-to-target and target-to-source transformations, respectively. The emotion classifier is used to specify the generator to generate target samples with specified emotion labels.

Specifically, we use the original multimodal emotion dataset to pretrain the emotion classifier EC_s, thus guiding the training direction of GAN. We use classification loss to guide the generator to learn the complementary and consistent semantic information of multimodal features and eliminate the heterogeneity between modal features. To synthesize new samples, we add an emotion state vector Z to the generator $G_{T \to S}$ as its input. During model training, each source data corresponds to a specific target data, and their emotion labels should be consistent. We define the loss function between source data and target data as shown in the following equation:

$$\mathcal{L}_C = \sum_k y_k \log \left(EC_s \left(G_{T \to S} \left(G_{S \to T} \left(S_k \right) \right), Z_k \right) \right)$$
$$+ \sum_k y_k \log \left(EC_s \left(G_{T \to S} \left(S_k, Z_k \right) \right) \right) \quad (1)$$

where $\mathcal{L}_C$ represents the loss between the data generated by the source-to-target and target-to-source generators and the true labels, and the loss between the data generated by the source-to-target generator and the true labels. k represents the sequence number of the sample, and Z_k is a one-hot encoded vector identical to the emotion label of the source sample S_k. Furthermore, to enable the generator to map the target sample to the emotion category specified by EC_s when it is transformed into the source sample by the generator ($G_{T \to S}$), we impose a constraint on the emotion state vectors as shown in the following equation:

$$\mathcal{L}_{RC} = \sum_i y_r \log \left(EC_s \left(G_{T \to S} \left(T_i \right), Z_r \right) \right) \quad (2)$$

where $\mathcal{L}_{RC}$ represents the classification loss between the generated data T_i and the real labels after passing through the target-to-source generator. Z_r, y_r is the one-hot encoding vector of emotion category r.

Futhermore, we assume that $G_{S \to T}$ is the generator for unimodal sources, and $D_{T,F}$ is the discriminator for unimodal targets. $G_{S \to T}$ performs an encoding operation (Enc) on the noise data to generate samples that conform to the distribution law of the real data. $D_{T,F}$ maps the input data to the target

domain through the decoding operation (Dec). We introduce discriminative loss to guide the model to learn intermodal complementary and consistent semantic information as shown in the following equation:

$$\mathcal{L} \left(G_{S \to T}, D_{T,F} \right)$$
$$= \mathbb{E}_{T \sim P_{\text{data}(T)}} \left[\log D_{T,F}(T) \right]$$
$$+ \mathbb{E}_{S \sim P_{\text{data}(T)}} \left[\log \left(1 - D_{T,F}(\text{Dec}(\text{Enc}(S))) \right) \right]. \quad (3)$$

Finally, to ensure the consistency of the distribution law of the generated data and the original data, this article also adds the identity loss in the training process of the model, and its loss function is defined as shown in the following equation:

$$\mathcal{L}_{\text{identity}} = \mathbb{E}_{S_i \sim P_{S_i}} \left[\left\| G_{T_i \to S_i} \left(S_i \right) - S_i \right\|^2 \right]$$
$$+ \mathbb{E}_{T_i \sim P_{T_i}} \left[\left\| G_{S_i \to T_i} \left(T_i \right) - T_i \right\|^2 \right] \quad (4)$$

where $\mathcal{L}_{\text{identity}}$ represents the square error between the data generated by the source data S_i through the target to the source generator and S_i, and the square error between the data generated by the target data T_i through the source to the target generator and T_i.

Therefore, the entire loss function $\mathcal{L}_{\text{EmoGAN}}$ of the generative adversarial network used in this article during the training process is defined as shown in the following equation:

$$\mathcal{L}_{\text{EmoGAN}} = \lambda_1 \mathcal{L}_{\text{identity}} + \lambda_2 \mathcal{L} \left(G_{S \to T}, D_{T,F} \right)$$
$$+ \lambda_3 \left(\mathcal{L}_C + \mathcal{L}_{RC} \right) \quad (5)$$

where λ_1, λ_2, λ_3 are the weights of $\mathcal{L}_{\text{identity}}$, $L(G_{S \to T}, D_{T,F})$ and $\mathcal{L}_C + \mathcal{L}_{RC}$ loss functions, which are learnable network parameters.

During model training, we use the Adam optimization algorithm to update the network parameters of the generator and discriminator. Among them, the update formula of the generator is defined as shown in the following equation:

$$g \leftarrow \frac{1}{k} \nabla_{\theta_D} \sum_{i=1}^k \mathcal{L} \left(G_{S \to T}, D_{T,F} \right). \quad (6)$$
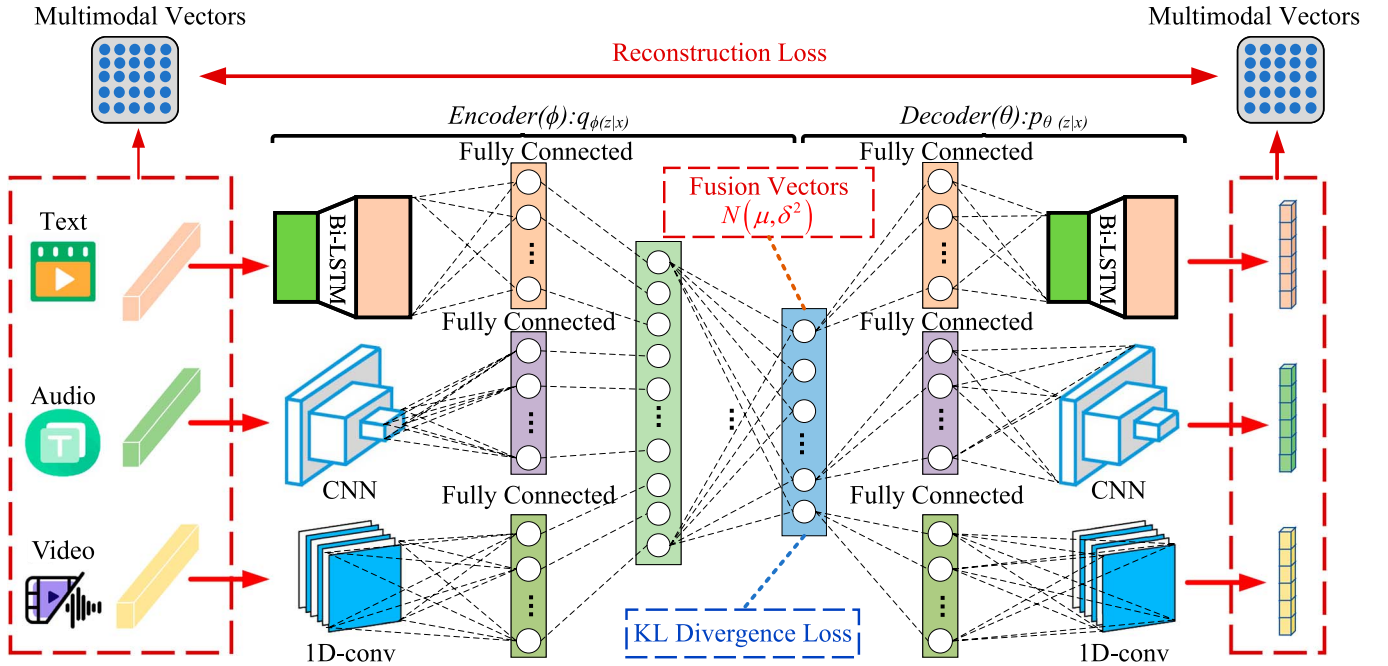
Fig. 4. Model architecture of DJVAE consists of an encoder and decoder. DJVAE includes Bi-LSTM for textual features, CNN for video, and 1D-conv for audio, where μ stands for mean, and δ stands for variance. In the model's training process, we use the KL divergence as the optimization function of the model.

Besides, the updated equation of the discriminator is defined as shown in the following equation:

$$g \leftarrow \frac{1}{k} \nabla_{\theta_D} \sum_{i=1}^{k} (\lambda_1 \mathcal{L}_{\text{identity}} + \lambda_2 \mathcal{L} (G_{S \to T}, D_{T,F}) \\ + \lambda_3 (\mathcal{L}_C + \mathcal{L}_{RC})). \tag{7}$$

After GAN is trained, this article uses it to generate multimodal emotional samples that conform to the original data distribution law for data augmentation. In particular, GAN networks are trained separately.

2) Intermodal Feature Fusion: To capture complementary semantic information between modalities and fuse multimodal feature vectors with differences, we design a DJVAE. As shown in Fig. 4, DJVAE consists of an encoder and a decoder. The encoder is used to map the data samples x into a low-dimensional feature space $z \in Z$, and the decoder is used to reconstruct the original data samples. The equation is defined as shown in the following equation:

$$\begin{cases} \text{DJVAE - Encoder: } \chi \to Z, \quad f(x) = z \\ \text{DJVAE - Decoder: } Z \to \chi, \quad g(z) = x' \end{cases}. \tag{8}$$

Then DJVAE obtains the optimal mapping relationship between the data samples and the low-dimensional feature space by minimizing the gap between the original data samples x and the reconstructed data samples. However, simple VAE cannot filter noise samples, but can only achieve point-to-point mapping between sample data and low-dimensional feature space through the mean square error (MSE Loss). Different from simple VAE, the DJVAE model proposed in this article will introduce KL divergence to estimate the similarity between the

encoder and decoder, so as to learn the latent semantic information of multimodal features. The formula for KL divergence is defined as shown in the following equation:

$$D_{KL} (q_\Phi(z \mid x) \| p_\vartheta(z \mid x)) = \sum_{z \in Z} q_\Phi(z \mid x) \log \frac{q_\Phi(z \mid x)}{p_\vartheta(z \mid x)} \\ = \sum_{z \in Z} q_\Phi(z \mid x) [\log q_\Phi(z \mid x) - \log p_\vartheta(z \mid x)] \\ = \sum_{z \in Z} q_\Phi(z \mid x) [\log q_\Phi(z \mid x) - \log p_\vartheta(x, z)] + \log p_\vartheta(x) \\ = \mathbb{E}_{q_{\Phi(Z|X)}} [\log q_\Phi(z \mid x) - \log p_\vartheta(z \mid x)] + \log p_\vartheta(z \mid x) \tag{9}$$

where $q_\Phi(z \mid x)$ represents the encoder's mapping of raw data samples to the latent feature space Z. $p_\vartheta(z \mid x)$ as an approximation of the true posterior distribution.

The above equation can be deformed to obtain the following equation:

$$\log p_\vartheta(x) = \mathbb{E}_{q_{\Phi(Z|X)}} [\log p_\vartheta(x, z) - \log q_\Phi(z \mid x)] \\ + D_{KL} (q_\Phi(z \mid x) \| p_\vartheta(z \mid x)). \tag{10}$$

At the same time, since the KL divergence is nonnegative, we can get the following equation:

$$\log p_\vartheta(x) \geq \mathbb{E}_{q_{\Phi(Z|X)}} [-\log q_\Phi(z \mid x) + \log p_\vartheta(x, z)] \\ = \mathbb{E}_{q_{\Phi(Z|X)}} [-\log q_\Phi(z \mid x) + \log p_\vartheta(x \mid z) \\ + \log p_\vartheta(x)] \\ = \mathbb{E}_{q_{\Phi(Z|X)}} \log p_\vartheta(x \mid z) - D_{KL} (q_\Phi(z \mid x) \| p_\vartheta(z)). \tag{11}$$

Therefore, we can get the loss function of DJVAE as shown in the following equation:

$$
\begin{aligned}
\text{Loss} &= -\mathbb{E}_{q_{\Phi(Z|X)}} \log p_{\vartheta}(x \mid z) \\
&\quad + D_{KL}\left(q_{\Phi}(z \mid x) \| p_{\vartheta}(z)\right)
\end{aligned} \tag{12}
$$

where $q_{\Phi}(z \mid x)$ represents the encoder's mapping of raw data samples to the latent feature space Z. $p_{\vartheta}(x \mid z)$ represents the decoder sampling data samples from the latent feature space Z. To simplify the calculation of KL divergence, we will use a standard normal distribution with mean 0 and variance 1.

3) Intramodal Context Feature Extraction: The utterances spoken by the speaker are arranged according to certain grammatical rules, and utterances composed of words in different sequences may have completely different meanings. In addition, such astext features, the feature vectors of the two modalities of video and audio contain the semantic information of the time dimension, and the speaker may show different emotions at different times. More importantly, the semantic information of the above three modalities is transmitted in a particular order. Therefore, we use Bi-LSTM for contextual feature extraction within the modality. Each LSTM block consists of multiple basic LSTM cells, and each LSTM cell contains an input gate, a forget gate, and an output gate [41].

The input gate is defined as shown in the following equation:

$$
\begin{aligned}
x_t &= \text{concat}\left([\alpha_w, \alpha_a, \alpha_v]\right) \\
i_t &= \text{sigmoid}\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)
\end{aligned} \tag{13}
$$

where $x_t \in \mathbb{R}^{d_f}$ is composed of three modal feature vectors of word vector α_ω, video feature vector α_v and audio feature vector α_a after feature fusion. i_t represents the input gate, which is used to process the input multimodal emotion feature vector. $W_i \in \mathbb{R}^{d_h \times d_k}$ is the weight matrix of the input gate, which is a learnable parameter. d_h is the number of units in the LSTM hidden layer, $d_k = d_f + d_h$, h_{t-1} represents the hidden layer state at time $t-1$. $b_i \in \mathbb{R}^{d_h}$ is the bias vector of the input gate.

Bi-LSTM is composed of the feature vector splicing of two hidden layers in opposite directions, and it is defined as shown in the following equation:

$$
\begin{aligned}
l_t &= \left[\overrightarrow{h_t} : \overleftarrow{h_t}\right] \\
V &= \text{concat}\left([l_1, l_2, \ldots, l_T]\right).
\end{aligned} \tag{14}
$$

Among them, $\overrightarrow{h_t}$ represents the forward hidden layer feature vectors, and $\overleftarrow{h_t}$ represents the reverse hidden layer feature vectors. l_t represents the hidden layer feature vectors at time t. V is composed of the concatenation of hidden layer feature vectors at all times.

4) Graph Interaction Network: We use graph to construct the interaction between speakers to capture the contextual semantic information related to the speakers. However, datasets in MERC have data imbalance issues, which will cause the model to fail to learn unbiased representations of minority class nodes, or even treat them as outliers in the data. Therefore, in response to the above problems, we propose a multitask graph neural network model, named MGNN, to alleviate the problem

of unbalanced distribution. MGNN simultaneously performs two subtasks to improve the generalization ability of GCN. 1) Data reconstruction; 2) emotion classification.

First, we construct a directed graph $G = \{V, \varepsilon, \mathcal{R}, W\}$, where the node $v_i (v_i \in V)$ is composed of multimodal feature vector g_i, and edge $r_{ij}(r_{ij} \in \varepsilon)$ is composed of the relationship between node v_i and node v_j, $\omega_{ij}(\omega_{ij} \in W, 0 \leq \omega_{ij} \leq 1)$ is the weight of the edge r_{ij}, and $r \in \mathcal{R}$ represents the relation type of the edge.

Edge Weights: Similarity attention mechanism is used to calculate the weights of edges in the graph, and aggregate neighbor information according to the calculated edge weights. We utilize multilayer perceptron (MLP) to calculate the similarity between node i and its surrounding neighbor j. The formula is defined as shown in the following equation:

$$
s_{ij}^{(t)} = W_{\theta_1}^{(t)}\left(\text{ReLU}\left(W_{\theta_2}^{(t)}\left[g_i^{(t-1)} \oplus g_j^{(t-1)} \oplus \Pi_{ij}\right]\right)\right). \tag{15}
$$

Among them, $W_{\theta_1}^{(t)}, W_{\theta_2}^{(t)}$ are the weight matrices of the tth layer in the multilayer perceptron network, which are learnable parameters. In the experiments, we set $W_{\theta_1}^{(t)}, W_{\theta_2}^{(t)}$ to 200 and 110, respectively. $\oplus$ represents the feature vector concatenation operation. $\prod_{ij} \in \{0, 1\}$, and $\prod_{ij} = 0$ means that there is no edge between node i and node j, and $\prod_{ij} = 1$ means that there is a directed edge between node i and node j. Next, we use the softmax function to get the attention score for each edge, as shown in the following equation:

$$
w_{ij}^{(t)} = \text{softmax}\left(s_{ij}^{(t)}\right) = \frac{\exp\left(s_{ij}^{(t)}\right)}{\sum_{n \in \mathcal{M}_i} \exp\left(s_{ij}^{(t)}\right)} \tag{16}
$$

where $\mathcal{M}_i$ is the set of surrounding neighbor nodes of node i. The larger $w_{ij}^{(t)}$ is, the closer the interaction between node i and node j is.

Message passing: Due to the serious data imbalance problem in MERC, if the GCN operation is used to aggregate the information of all surrounding neighbor nodes, it will cause the model to be biased toward fitting the majority class nodes, while the minority class nodes are regarded as outliers in the data. Therefore, we consider it unnecessary to aggregate all neighbor nodes in the graph. As shown in Fig. 5, to solve the above problems, we randomly perform mask operation on some neighbors, then use graph convolution operation to aggregate information for neighbor nodes that have not been masked, and then perform data reconstruction tasks. The message passing is defined as shown in the following equation:

$$
\phi_i^{(t)} = \text{ReLU}\left(\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{M}_i^r} \frac{w_{ij}}{c_{i,r}} W_r^{(t)} \phi_j^{(t-1)} + w_{i,i} W_\zeta^{(t)} \phi_i^{(t-1)}\right) \tag{17}
$$

where $\mathcal{M}_i^r$ is the set of unmasked neighbors around node i under edge relation $r \in \mathcal{R}$, and w_{ij} is the attention score between node i and node j under edge relation $r \in \mathcal{R}$. $c_{i,r}$ is the size of the modulus of M_i^r. $W_r^{(t)}, W_\zeta^{(t)}$ are learnable weight matrices. In the experiments, we set $W_r^{(t)}$ and $W_\zeta^{(t)}$ to 150
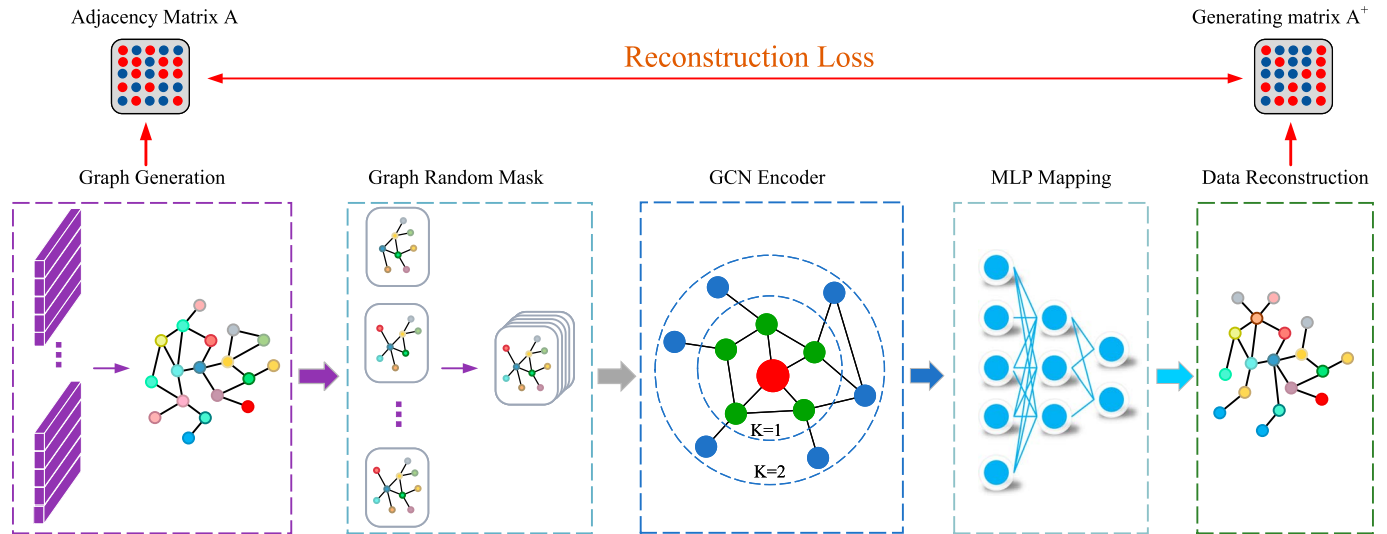
Fig. 5. Model architecture for graph masking and data reconstruction consists of graph structure generation, graph random masking, GCN encoder, and MLP mapping. First, we randomly mask the generated graph nodes, and then use GCN to aggregate the information of first-order neighbors and second-order neighbors. Finally, MLP is used for data reconstruction. In the data reconstruction process, we use MSE Loss to measure the gap between the original and predicted data.

and 400, respectively. $\phi_i^{(t)}$ represents the feature vector containing speaker information aggregated after convolution operation. $\phi_i^{(\bar{t})}$ represents the multimodal feature vectors aggregated after graph convolution operation.

Data reconstruction: We use MSE Loss to measure the difference between the reconstructed data and the original data. The equation is defined as shown in the following equation:

$$L_{\text{recon}} = \sum_{i=1}^{N} \frac{1}{N} \left(y_i - \hat{y}_i\right)^2 \quad (18)$$

where y_i represents the multimodal feature vector of the original data, and $\hat{y}_i$ represents the predicted multimodal feature vector. In general, the smaller L_{recon}, the stronger the ability of model data reconstruction.

We will introduce the specific implementation process of the sentiment classification task in detail in the next section.

5) Emotion Classification: In model training, compared with the majority class samples, the minority class samples have little influence on the model, which will cause the model to update the parameters in a direction that is beneficial to the majority class samples. However, it is also necessary for the model to be able to correctly classify the minority class samples. Therefore, we use an ensemble learning algorithm called Adaboost, which continuously optimizes the weight of utterance samples in the weak classifier and increases the weight of the minority class utterance samples in the classification process, thereby forming a strong classifier.

B. Model Training

The CBERL model makes the model pay more attention to indistinguishable samples by adding a conditioning factor $(1 - P_{i,j}[y_{i,j}])^\gamma$ to the cross-entropy loss function, and the $L2$ regularization method is used to prevent the model from overfitting, thereby providing guidance for guiding the updated direction of the model parameters. The loss is defined as shown in the following equation:

$$L_{\text{focal}} = -\frac{1}{\sum_{\xi=1}^{N} c(\xi)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} \left(1 - P_{i,j}\left[y_{i,j}\right]\right)^\gamma \log \mathrm{P}_{i,j}\left[y_{i,j}\right] + \lambda \|\theta\|_2 \quad (19)$$

where N is the number of samples in the benchmark dataset, $c(i)$ represents the number of utterances contained in sample i, and $P_{i,j}$ is the probability distribution of the emotion category in the jth sentence in the ith dialogue, which is a number in the range $[0, 1]$. γ is a constant greater than 0, λ is the weight decay coefficient, and θ is the set of all learnable parameters in the network.

Then we can get the total loss function for model training as shown in the following equation:

$$L_{\text{classify}} = \xi_1 L_{\text{recon}} + \xi_2 L_{\text{focal}} . \quad (20)$$

Among them, ξ_1, ξ_2 represent the importance of data reconstruction and emotion classification tasks, respectively. In general, the smaller the loss function value, the better the effect of the model's emotion classification.

IV. EXPERIMENTS

A. Benchmark Dataset Used

The MELD [42] and IEMOCAP [43] benchmark datasets are widely used by researchers in MERC research. Since the IEMOCAP benchmark dataset does not provide a separate validation set, we use 10% of the training data as the validation set for the model. Furthermore, in these two benchmark datasets, the ratio of training and testing sets is 80:20. In particular, the first four sessions of the dataset are used for training and the last session of the dataset is used for testing in the IEMOCAP dataset.

TABLE I
COMPARED WITH OTHER BASELINE MODELS ON THE IEMOCAP DATASET

| Methods | IEMOCAP | | | | | | | | | | | | | |
| | Happy | | Sad | | Neutral | | Angry | | Excited | | Frustrated | | Average(w) | |
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	WAA	WAF1
TextCNN	27.73	29.81	57.14	53.83	34.36	40.13	61.12	52.47	46.11	50.09	62.94	55.78	48.93	48.17
bc-LSTM	29.16	34.49	57.14	60.81	54.19	51.80	57.03	56.75	51.17	57.98	67.12	58.97	55.23	54.98
DialogueRNN	25.63	33.11	75.14	78.85	58.56	59.24	64.76	65.23	80.27	71.85	61.16	58.97	63.42	62.74
DialogueGCN	40.63	42.71	**89.14**	**84.45**	61.97	63.54	67.51	64.14	65.46	63.08	64.13	66.90	65.21	64.14
CT-Net	47.97	51.36	78.01	79.94	**69.08**	65.82	72.98	67.21	85.35	**78.74**	52.27	58.83	68.01	67.55
LR-GCN	54.24	55.51	81.67	79.14	59.13	63.84	69.47	69.02	76.37	74.05	68.26	68.91	68.52	68.35
EmoBERTa	56.94	56.43	79.19	83.02	64.04	61.56	70.68	69.64	**86.05**	78.09	63.84	68.78	67.38	67.37
M2FNet	55.92	58.79	69.18	71.97	65.80	**66.31**	75.31	70.76	74.84	73.15	66.87	67.01	67.64	68.06
DialogXL	**61.18**	64.81	77.18	79.15	64.84	58.35	65.48	65.48	77.97	77.97	55.28	58.02	66.05	65.38
CBERL	58.84	**67.34**	63.31	72.84	56.42	60.75	**75.32**	**73.51**	70.32	70.77	**78.21**	**71.19**	**69.36**	**69.27**

Note: Acc. represents the accuracy. Average(w) represents the weighted average. The best result in each column is in bold.

B. Evaluation Metrics

In this section, we use the following four metrics to evaluate the performance of dialogue emotion recognition on the IEMOCAP and MELD benchmark datasets: 1) accuracy; 2) weighted average accuracy (WAA); 3) F1-score; 4) weighted average F1-score (WAF1).

C. Baseline Models

To validate the effectiveness of CBERL on MERC, the article compared the following baseline models with our model: TextCNN [44], bc-LSTM [45], DialogueRNN [1], DialogueGCN [28], CT-Net [19], LR-GCN [46], EmoBERTa [47], M2FNet [48], and DialogXL [49]. We use the data from the original article as our comparative experiment.

V. RESULTS AND DISCUSSION

A. Comparison with Baseline Methods

We compare the proposed multimodal emotion recognition in conversations method CBERL with current baseline models. Tables I and II show the recognition accuracy and F1 value of each category of CBERL and other baseline models on the IEMOCAP and MELD benchmark datasets, respectively.

IEMOCAP: As shown in Table I, compared with other baseline models, the CBERL has the best overall recognition performance on the IEMOCAP benchmark dataset, and the WAA and WAF1 values are 69.3% and 69.2%, respectively. CBERL proposes a method that combines data augmentation, multimodal feature fusion, and interaction between speakers for emotion recognition. Among other comparison algorithms, the effect of LR-GCN is second, with WAA and WAF1 values of 68.5% and 68.3%, respectively. We believe this is because LR-GCN comprehensively considers the interaction between speakers as well as the latent relationship between utterances. The four methods CT-Net, EmoBERTa, M2FNet, and DialogXL are all based on the Transformer architecture to achieve multimodal emotion recognition, and their emotion recognition accuracy rates are relatively close and better than bc-LSTM. The performance improvement may be attributed to the fact that Transformer models long-term contextual correlations better than LSTM. The emotion recognition effect of DialogueRNN and DialogueGCN is slightly worse than that of CBERL and LR-GCN, with WAA

being 63.4% and 62.7%, respectively, and WAF1 being 62.7% and 64.1%, respectively. We think this is because DialogueRNN and DialogueGCN do not exploit the complementary semantic information between modalities. The emotion recognition effect of TextCNN and bc-LSTM is very poor, WAA is 20.4% ~ 14.1% lower than other baseline models and CBERL, and WAF1 is 21.1% ~ 14.3% lower than other baseline models and CBERL. We guess this is because TextCNN and bc-LSTM ignore the interaction between speakers and the emotional fusion between modalities.

MELD: As shown in Table II, CBERL has the best emotion recognition performance among all comparison algorithms, with WAA and WAF1 values of 67.7% and 66.8%, respectively. Specifically, our proposed model CBERL significantly improves emotion recognition performance on two minority classes "fear" and "disgust" labels. On the "fear" emotion label, CBERL achieves 25.0% and 22.2% values on WAA and WAF1, respectively. On the "disgust" emotion label, CBERL achieves 25.8% and 24.6% values on WAA and WAF1, respectively. Compared with other comparison algorithms, CBERL is 1% to 11.4% more effective on WAA and 1.2% to 11.8% more effective on WAF1. We believe that the significant improvement of CBERL on minority emotion can be attributed to the following four aspects:

1) Increase the proportion of minority labels in all emotion categories by using data augmentation methods, which provides a data basis for subsequent models to learn discriminative class boundaries.
2) The two tasks of data reconstruction and emotion prediction are performed in parallel by randomly masking the graph nodes. CBERL can improve the model's ability to learn unbiased representations of minority class nodes.
3) To encourage the model to focus on indistinguishable classes by adding an adjustment factor to the cross-entropy loss function. CBERL can assign higher loss weights to minority class labels.
4) Improve the accuracy of emotion classification results by synthesizing the classification results of multiple weak classifiers.

The above advantages significantly improve the performance of CBERL on the MELD benchmark dataset, especially on the minority class emotion labels.

TABLE II
COMPARED WITH OTHER BASELINE MODELS ON THE MELD DATASET

| Methods | MELD | | | | | | | | | | | | | | | |
| | Neutral | | Surprise | | Fear | | Sadness | | Joy | | Disgust | | Anger | | Average(w) | |
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	WAA	WAF1
TextCNN	76.23	74.91	43.35	45.51	4.63	3.71	18.25	21.17	46.14	49.47	8.91	8.36	35.33	34.51	56.35	55.01
bc-LSTM	78.45	73.84	46.82	47.71	3.84	5.46	22.47	25.19	51.61	51.34	4.31	5.23	36.71	38.44	57.51	55.94
DialogueRNN	72.12	73.54	54.42	49.47	1.61	1.23	23.97	23.83	52.01	50.74	1.52	1.73	41.01	41.54	56.12	55.97
CT-Net	75.61	77.45	51.32	52.76	5.14	10.09	30.91	32.56	54.31	56.08	11.62	11.27	42.51	44.65	61.93	60.57
LR-GCN	**81.53**	80.81	55.45	57.16	0.00	0.00	36.33	36.96	62.21	**65.87**	7.33	11.01	52.64	54.74	66.71	65.67
EmoBERTa	78.92	82.53	50.22	50.21	1.87	1.92	33.31	31.22	72.15	61.77	9.11	2.53	43.36	46.48	64.15	65.24
M2FNet	72.88	67.98	62.76	**58.35**	5.57	2.93	50.09	**48.24**	68.49	64.92	17.69	15.79	**57.33**	**55.33**	66.74	66.12
DialogueXL	65.29	77.57	**80.00**	53.21	0.00	0.00	66.67	43.72	**68.98**	55.92	0.00	0.00	53.67	51.06	65.10	63.31
CBERL	81.45	**82.03**	55.24	57.91	**25.04**	**22.23**	47.51	41.36	66.03	65.67	**25.81**	**24.65**	53.75	55.31	**67.78**	**66.89**

Note: Acc. represents the accuracy. Average(w) represents the weighted average. The best result in each column is in bold.
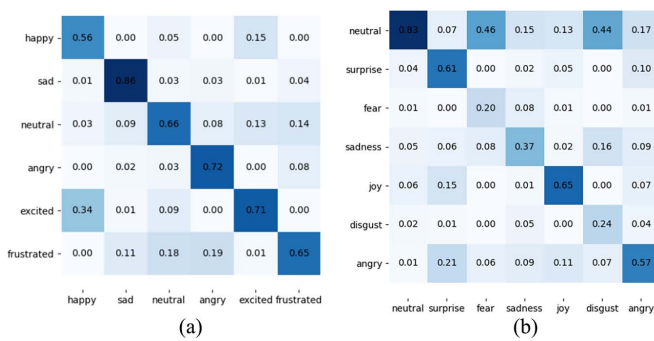

Fig. 6. Confusion matrix learned by CBERL on the IEMOCAP and MELD benchmark datasets. (a) Confusion matrix obtained by CBERLon the IEMO-CAP dataset. (b) Confusion matrix obtained by CBERLon the MELD dataset.

By conducting extensive experiments, we demonstrate that CBERL can effectively capture contextual information within modalities, complementary semantic information between modalities, and interactions between speakers. CBERL can effectively utilize this semantic information to improve the model's emotion classification ability.

B. Analysis of the Experimental Results

To give the specific classification of CBERL on the benchmark datasets, we show the confusion matrices obtained by CBERL on the IEMOCAP and MELD benchmark datasets in Fig. 6. On the IEMOCAP benchmark dataset, the semantic information learned by the model between the two emotions is similar due to the small discrimination between the emotions "happy" and "excited". Therefore, the model is prone to confusion about these two emotions. We can also observe the confusion matrix and find that the model is apt to misclassify the "happy" label as the "excited" label, and the "excited" class as the "happy" class. For the "sad" class emotion, the model has difficulty distinguishing it from the "frustrated" class emotion. In the trained multimodal corpus, all emotion categories have some relationship with the "neutral" emotion label so that the model may misclassify the "neutral" emotion as other categories and vice versa. Someone with an "angry" emotion may usually be accompanied by a "frustrated" emotion. Therefore,

the model may learn this semantic information during training, causing the model to incorrectly classify the "angry" emotion as the "frustrated" emotion. We observed the confusion matrix on the MELD benchmark dataset and found that the model's recognition accuracy on the "fear" and "disgust" emotion labels improved significantly. Compared with other baseline models that barely recognize "fear" and "disgust" emotions, we believe this is mainly due to our increased amount of data for "fear" and "disgust" emotions, which effectively alleviates the imbalance in data distribution. The number of utterances belonging to the "neutral" emotion is the largest among all emotion categories, which makes the model biased toward learning feature representations for utterances with "neutral" emotion, which makes it easy for the model to misclassify other emotions as "neutral" emotion.

C. Visualization

To compare the distribution of different emotions in the feature space more intuitively, we visualized the original multimodal emotion features and the emotion features learned by different networks from the IEMOCAP benchmark dataset. Specifically, we project the original emotion features as well as the learned high-dimensional emotion features into a two-dimensional space. In this article, we use the t-SNE method [50] to visualize emotion features in the IEMOCAP dataset and color each point according to the emotion label.

As shown in Fig. 7(a), we find that the distribution of the original multimodal data without network processing in the 2-D space is very messy, there is no distribution pattern among the various emotion categories, which are mutually fused. As shown in Fig. 7(b) and 7(c), we find that the embedding representations learned by the bc-LSTM and DialogueRNN models are much better than the original data distribution. There is a boundary between utterances belonging to different emotion categories. However, since neither the bc-LSTM nor the DialogueRNN model considers the feature fusion between modalities and the interaction between speakers, the distinguishable boundaries are still blurred. As shown in Fig. 7(d), the visualization of the embedding representation learned by CBERL is the best. After considering the three influencing factors of modeling
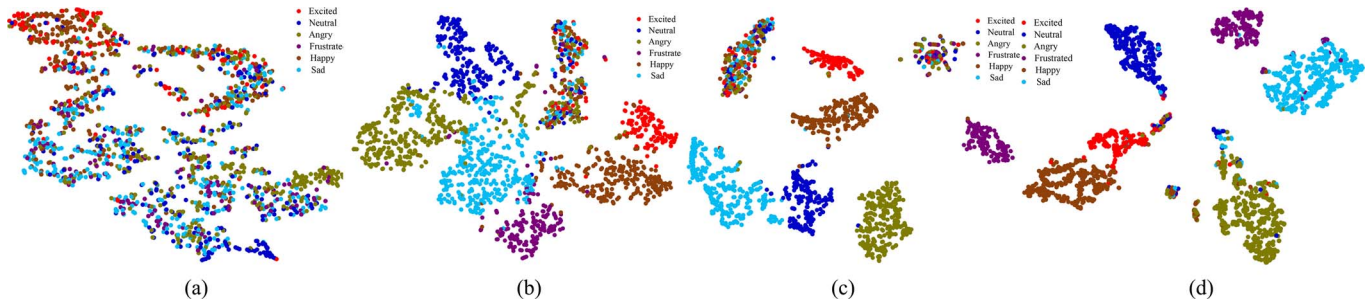
Fig. 7. Visualizing feature embeddings for the multimodal emotion on the IEMOCAP benchmark dataset. Each dot represents an utterance, and its color represents an emotion. (a) Raw data distribution. (b) Learned by bc-LSTM. (c) Learned by DialogueRNN. (d) Learned by CBERL.

TABLE III
CBERL PERFORMS ABLATION EXPERIMENTS ON THE IEMOCAP
BENCHMARK DATASET

Feature Fusion	γ	Graph Node Mask	Adaboost	WAF1
+	-	-	-	67.1
-	+	-	-	66.0
-	-	+	-	66.8
-	-	-	+	65.9
+	+	-	-	67.9
+	-	+	-	68.5
+	-	-	+	68.0
-	+	+	-	67.4
-	+	-	+	66.2
-	-	+	+	68.3
+	+	+	-	68.8
+	-	+	+	69.0
+	+	-	+	68.2
-	+	+	+	67.9
-	-	-	-	64.1
+	+	+	+	69.2

intramodal contextual semantic information, fusion of semantic information between modalities, and emotional interactions between speakers, CBERL learns embedding representations with high intraclass similarity and interclass variability, and different emotional labels are partitioned by different boundaries between them.

D. Ablation Study

In this section, we perform ablation experiments on each part of our proposed model CBERL on the IEMOCAP benchmark dataset. The results are analyzed to judge the impact of each module of CBERL on the effect of emotion recognition. The specific results of the ablation experiments are shown in Table III. In particular, when using the adjustment factor γ, we set $\gamma = 3$.

We conducted a total of 16 groups of ablation experiments to compare the effects of the algorithms. When only one module is chosen as a component of CBERL, we find that feature fusion has the greatest impact on the model, and the WAF1 value of emotion recognition can reach 67.1%. The effect of the graph node mask is second, and the WAF1 value is 65.9%. Furthermore, as shown in Table I, the WAF1 value of DialogueGCN is 64.1%, and the experimental results demonstrate the effectiveness of graph node masks. The effects of γ and the Adaboost algorithm are relatively similar, with WAF1 values of

66.0% and 65.9%, respectively, which are lower than the effects of feature fusion and graph node masking. When selecting two modules as components of CBERL, the combination of feature fusion and graph node mask has the best effect on emotion recognition, with a WAF1 value of 68.5%. The combination of γ and the Adaboost algorithm has the lowest effect, with a WAF1 value of only 66.2%, but slightly higher than only γ or the Adaboost algorithm. The emotion recognition effects of other composition methods are similar and are all higher than the emotion recognition effects of only a single module. When three modules are selected as the components of CBERL, the combination of feature fusion, graph node mask, and the Adaboost algorithm performs the best emotion recognition with a WAF1 value of 69.0%. The emotion recognition effect of other composition methods is also higher than that of CBERL composed of two modules. When choosing four modules as components of CBERL, emotion recognition performed best in all ablation experiments, with a WAF1 value of 69.2%. When none of the modules is selected (i.e. just Bi-LSTM and GCN modules with data augmentation) as a component of CBERL, the model has the worst emotion recognition performance, and the WAF1 value of emotion recognition is 64.1%. Experiments demonstrate the effectiveness of each module. In particular, when not using the Adaboost algorithm as our emotion classifier, we use a multilayer perceptron (MLP) as our emotion classifier.

E. Analysis on Parameters

In this section, we analyze the effect of the conditioning factor γ in the cross-entropy loss function on the model performance. γ is an important hyperparameter of CBERL, which can adjust the weight of indistinguishable samples in the loss function, so that the model can focus on the classification of indistinguishable samples. Therefore, we designed several sets of comparative experiments to select the best γ value. Specifically, we choose γ values from the set $M = \{0, 1, 2, 3, 4, 5\}$. The experimental results obtained by our different values of γ on the IEMOCAP benchmark dataset are shown in Fig. 8. As the value of γ increases from 0, CBERL achieves a certain degree of performance improvement. When $\gamma = 3$, the model achieves the best performance. The best WAA value of CBERL can reach 69.3%, and the WAF1 value can reach 69.2%. However, when $\gamma > 3$, the WAA value of CBERL will start to decrease.
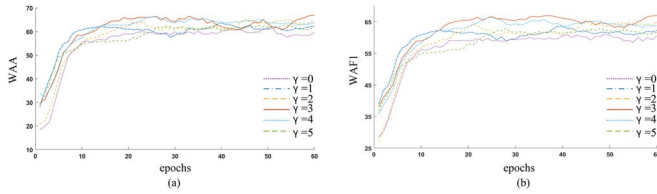
Fig. 8. We set different γ values to test their experimental effects on the IEMOCAP benchmark dataset. To make the results more intuitive, we perform Gaussian smoothing on the WAA and WAF1 values. (a) Weighted accuracy. (b) Weighted f1.
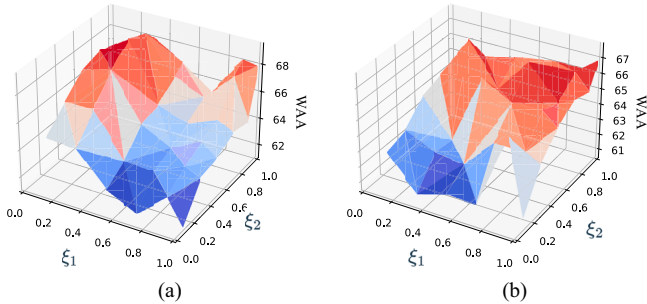

Fig. 9. We perform ablation studies to explore the emotion recognition accuracy of different values of ξ_1 and ξ_2 on the (a) IEMOCAP and (b) MELD datasets. Left: IEMOCAP dataset. Right: MELD datasets.

It indicates that too much concentration of the model on indistinguishable samples will make the model overfitting effect and will bring some redundant semantic information to the model, thus making it less effective in emotion recognition.

In addition, we also study the impact of parameters ξ_1 and ξ_2 on the experimental results. As shown in Fig. 9, in the IEMOCAP and MELD data sets, we use 3-D grid search to find the optimal values of parameter ξ_1 and ξ_2. Experimental results show that when the parameter range is between 0 and 1, the model's emotion recognition accuracy is greatly affected by the parameters, and when $\xi_1 = 0.3$ and $\xi_2 = 0.7$, the model can achieve the best emotion recognition effect.

F. Effectiveness of Data Augmentation

To verify the impact of the data augmentation module on the experimental results, we performed an ablation experiment on the data augmentation module. As shown in Table IV, We find that CBERL with a data augmentation module outperforms CBERL without a data augmentation module in emotion recognition on both IEMOCAP and MELD datasets. In addition, we also explore the experimental effects of some other baseline models after using the proposed data augmentation module. Experimental results show that the emotion recognition effect of the baseline model using the data augmentation module can be significantly improved. The performance improvement may be attributed to the data augmentation module can balance the data distribution relationship among different emotion categories, which is beneficial to enhance the representation learning of graphs.

TABLE IV
EXPERIMENTAL RESULTS OF THE CBERL METHOD FOR MULTIMODAL EMOTION RECOGNITION TASKS ON THE IEMOCAP AND MELD DATASETS

Methods	IEMOCAP		MELD	
	WAA	WAF1	WAA	WAF1
TextCNN (w/o D)	48.93	48.17	56.35	55.01
TextCNN (D)	52.47	53.15	57.77	57.06
bc-LSTM (w/o D)	55.23	54.98	57.51	55.94
bc-LSTM (D)	59.73	59.17	58.33	57.69
DialogueRNN (w/o D)	63.42	62.74	56.12	55.97
DialogueRNN (D)	65.14	65.11	58.42	58.07
DialogueGCN (w/o D)	65.21	64.14	60.17	59.86
DialogueGCN (D)	66.59	66.17	61.24	61.37
CBERL (w/o D)	68.69	68.41	66.23	65.01
CBERL (D)	69.36	69.27	67.78	66.89

Note: CBERL (D) indicates that data augmentation is used.

TABLE V
EXPERIMENTAL RESULTS OF THE CBERL METHOD FOR MINORITY EMOTION RECOGNITION TASKS (I.E., HAPPY, FEAR, AND DISGUST) ON THE IEMOCAP AND MELD DATASETS

Methods	IEMOCAP	MELD	
	Happy	Fear	Disgust
CBERL (N)	47.48	3.15	2.96
CBERL (D)	65.34	18.92	21.36
CBERL (M)	61.19	13.40	17.71
CBERL (A)	58.62	10.77	12.52
CBERL	67.34	22.23	24.65

Note: WAF1 is chosen as the evaluation criterion for the experiments. CBERL (N) indicates that no data augmentation, masking strategy and adjustment factor are used. CBERL (D) indicates that only data augmentation is used without masking strategies and adjustment factors. CBERL (M) indicates that only masking strategies are used without data augmentation and adjustment factors. CBERL (A) indicates that only adjustment factors is used without masking strategies and data augmentation.

To verify whether the data augmentation module and the weighted cross entropy/masking strategy can improve the emotion recognition effect of minority emotions, we performed the ablation experiment of the data augmentation module and the weighted cross entropy/masking strategy on minority emotions (i.e., happy, fear, and disgust). As shown in Table V, the emotion recognition effect of CBERL (N) on minority emotions is particularly poor, especially on fear and disgust. The emotion recognition effect of CBERL (A) on minority emotions is slightly improved compared with CBERL (N). The improved performance may be attributed to the adjustment factor forcing the model to focus on the classification of hard samples. The emotion recognition effect of CBERL (M) is better than that of CBERL (N) on minority emotions. We speculate that the masking mechanism can alleviate the long-tail problem of graph node distribution while obtaining better node representation. Compared with CBERL (N), CBERL (M), and CBERL (A), CBERL (D) achieves the best emotion recognition performance on the minority class emotion. The performance improvement is attributed to the data augmentation module can optimize the data distribution of minority class emotion, which enables the model to learn better
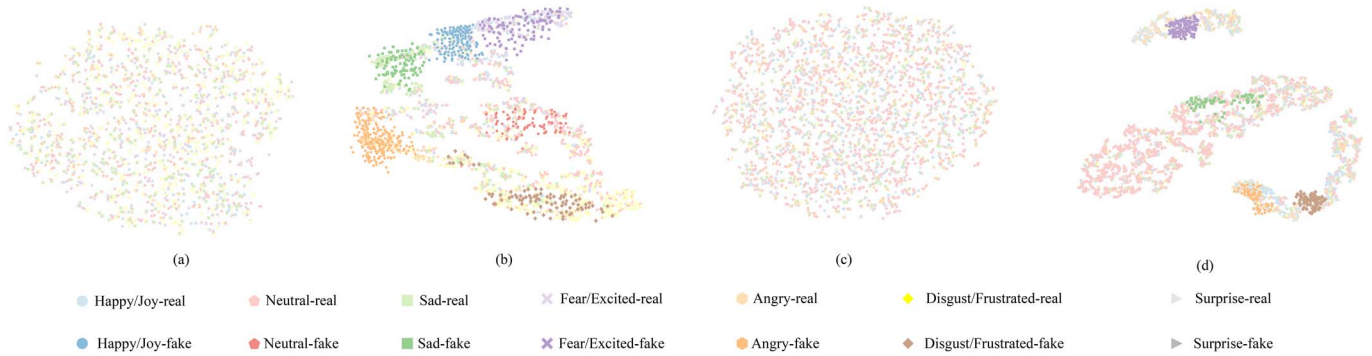
Fig. 10. Original data distribution and the data distribution after data augmentation of the IEMOCAP and MELD datasets. (a) Original distribution of the IEMOCAP dataset. (b) Distribution of IEMOCAP datase tafter data augmentation. (c) Original distribution of the MELD dataset. (d) Distribution of MELD dataset after data augmentation.

TABLE VI
EXPERIMENTAL RESULTS OF THE PROPOSED GAN METHOD AND OTHER
DATA AUGMENTATION METHODS ON MULTIMODAL EMOTION
RECOGNITION TASKS ON IEMOCAP AND MELD DATASETS

Methods	IEMOCAP		MELD	
	WAA	WAF1	WAA	WAF1
SMOTE	68.93	68.67	66.31	65.38
Deep SMOTE	68.98	68.86	66.57	65.43
GHM	69.12	69.05	66.89	66.14
GAN	68.64	68.31	66.17	66.00
CycleGAN	69.20	69.13	66.95	66.27
Proposed GAN	69.36	69.27	67.78	66.89

TABLE VII
DISTRIBUTION OF SAMPLES GENERATED USING GAN FOR
IEMOCAP AND MELD DATASETS

Categories	IEMCOCAP	MELD
Happy/Joy	250	0
Neutral	80	0
Sad	120	60
Fear/Excited	120	200
Anger	250	60
Disgust/Frustrated	80	200
Surprise	–	0

emotion class boundaries. CBERL performs best in emotion recognition in all experiments because it combines the advantages of data augmentation modules, masking strategies, and adjustment factors.

G. Comparison of Different Data Augmentation Methods

We conducted comparative experiments to illustrate the superiority of the data augmentation method GAN proposed in this article. As shown in Table VI, we compare the proposed GAN method with five data augmentation methods: SMOTE [38], Deep SMOTE [39], GHM [10], GAN [51], and CycleGAN [52]. Experimental results show that the proposed GAN method for data augmentation achieves the best emotion recognition results on the IEMOCAP and MELD data sets. The emotion recognition effect of CycleGAN is second, and the emotion recognition effect of GHM is slightly lower than CycleGAN. The performance improvement of SMOTE, Deep SMOTE, and GAN for emotion recognition is not obvious. Compared with the undersampling and oversampling data augmentation methods, the performance improvement may be attributed to the fact that the proposed GAN method can generate samples that conform to the original data distribution and expand the sample size of the dataset. Compared with the GAN and CycleGAN methods, the proposed GAN method can learn the complementary and consistent semantic information between multimodal features and achieve better data generation.

H. Emotion Distribution After Data Augmentation

As shown in Table VII, we give the number of samples increased for each emotion category for the IEMOCAP and MELD datasets. To eliminate the long-tail problem, we try to keep the number of each emotion category as consistent as possible. However, the number of some emotion categories is too small, which leads to adding too much data to fail to increase the performance of the model. In addition, we also show the spatial distribution of different emotion categories after data augmentation. As shown in Fig. 10, the original distributions of the IEMOCAP and MELD datasets are chaotic and indistinguishable, while the distributions between different emotion categories after data augmentation are discriminative. The data distribution enhanced by GAN can enhance the emotion recognition ability of subsequent models.

VI. CONCLUSION AND FUTURE WORK

This article proposes the CBERL model, a MERC framework for dialogue emotion recognition tasks. CBERL extracts intramodal contextual semantic information and fuses intermodal complementary information. Meanwhile, CBERL also considers the problem of data imbalance, and solves this problem from three levels of data augmentation, sampling strategy, and loss-sensitive. Extensive experiments are conducted on IEMOCAP and MELD datasets, and compared with other models, CBERL achieves better classification accuracy. Furthermore, we demonstrate the necessity of feature fusion and address the data imbalance problem.

In future work, in addition to considering the multimodal feature fusion and the imbalanced data distribution of emotion categories, we will also consider using the heterogeneous information of multimodal data to improve the emotion recognition accuracy of the model.

REFERENCES

[1] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 6818–6825.

[2] Y. Shou, T. Meng, W. Ai, S. Yang, and K. Li, "Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis," *Neurocomputing*, vol. 501, no. 3, pp. 629–639, 2022.

[3] A. Zadeh and P. Pu, "Multimodal language analysis in the wild: CMU-mosei dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers) (ACL)*, 2018, pp. 2236–2246.

[4] J. Liu et al., "Multimodal emotion recognition with capsule graph convolutional based representation fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Piscataway, NJ, USA: IEEE Press, 2021, pp. 6339–6343.

[5] N.-T. Tran, V.-H. Tran, N.-B. Nguyen, T.-K. Nguyen, and N.-M. Cheung, "On data augmentation for gan training," *IEEE Trans. Image Process.*, vol. 30, pp. 1882–1897, 2021.

[6] P. Dai, Y. Li, H. Zhang, J. Li, and X. Cao, "Accurate scene text detection via scale-aware data augmentation and shape similarity constraint," *IEEE Trans. Multimedia*, vol. 24, pp. 1883–1895, 2022.

[7] H. Chen, B. Du, S. Luo, and W. Hu, "Deep point set resampling via gradient fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2913-2930, Mar. 2023.

[8] H. Liu, J. Lin, S. Xu, T. Bi, and Y. Lao, "A resampling method based on filter designed by window function considering frequency aliasing," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 67, no. 12, pp. 5018–5030, Dec. 2020.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.* Piscataway, NJ, USA: IEEE Press, 2017, pp. 2980–2988.

[10] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 8577–8584.

[11] B.-H. Su and C.-C. Lee, "Unsupervised cross-corpus speech emotion recognition using a multi-source cycle-GAN," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1991–2004, Jul./Sep. 2023.

[12] A. Antoniou, A. Storkey, and H. Edwards, "Augmenting image classifiers using data augmentation generative adversarial networks," in *Proc. Artif. Neural Netw. Mach. Learn. (ICANN)*, Springer, 2018, pp. 594–603.

[13] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, "Emotion classification with data augmentation using generative adversarial networks," in *Proc. Adv. Knowl. Discovery Data Mining: 22nd Pacific-Asia Conf. (PAKDD)*, Melbourne, VIC, Australia: Springer, 2018, pp. 349–360.

[14] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Augmenting generative adversarial networks for speech emotion recognition," in *Proc. 2020 Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. International Speech Communication Association, 2020, pp. 521–525.

[15] G. Bao et al., "Data augmentation for EEG-based emotion recognition using generative adversarial networks," *Frontiers Comput. Neurosci.*, vol. 15, no. 9, p. 723843, 2021.

[16] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. MIT Press, 2017.

[17] T. Meng, Y. Shou, W. Ai, J. Du, H. Liu, and K. Li, "A multi-message passing framework based on heterogeneous graphs in conversational emotion recognition," *Available at SSRN 4353605*, 2021.

[18] X. Du et al., "An efficient LSTM network for emotion recognition from multichannel EEG signals," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1528-1540, Jul./Sep. 2022.

[19] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 985–1000, 2021.

[20] T. Zhang, X. Wang, X. Xu, and C. L. P. Chen, "Gcb-net: Graph convolutional broad network and its application in emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 379–388, Jan./Mar. 2022.

[21] D. Nguyen et al., "Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 1313–1324, 2022.

[22] L. Qin, W. Che, Y. Li, M. Ni, and T. Liu, "DCR-Net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 05, 2020, pp. 8665–8672.

[23] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3504–3514, 2021.

[24] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Syst. Appl.*, p. 121692, 2023.

[25] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, "A transformer-based model with self-distillation for multimodal emotion recognition in conversations," *IEEE Trans. Multimedia*, vol. 26, pp. 776–788, 2023.

[26] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1426–1439, Jul./Sep. 2020.

[27] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2122–2132.

[28] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proc. Conf. Empirical Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*. 2019, pp. 154–164.

[29] L. Yi and M.-W. Mak, "Improving speech emotion recognition with adversarial data augmentation network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 172–184, Jan. 2022.

[30] S. Zeng, B. Zhang, J. Gou, and Y. Xu, "Regularization on augmented data to diversify sparse representation for robust image classification," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4935–4948, Jun. 2022.

[31] S. Wang, Y. Yang, Z. Wu, Y. Qian, and K. Yu, "Data augmentation using deep generative models for embedding based speaker recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2598–2609, 2020.

[32] J.-S. Kang, S. Kavuri, and M. Lee, "ICA-evolution based data augmentation with ensemble deep neural networks using time and frequency kernels for emotion recognition from EEG-data," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 616–627, Sep. 2019.

[33] X. Zhang et al., "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4386–4399, Sep. 2021.

[34] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Volume 1 Long Papers)*. 2018, pp. 2247–2256.

[35] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Nat. Lang. Process.* 2017, pp. 1103–1114.

[36] W. Zhou, X. Lin, J. Lei, L. Yu, and J.-N. Hwang, "MFFENet: Multiscale feature fusion and enhancement network for RGB–thermal urban road scene parsing," *IEEE Trans. Multimedia*, vol. 24, pp. 2526–2538, 2022.

[37] H. Yu, C. Sun, X. Yang, S. Zheng, and H. Zou, "Fuzzy support vector machine with relative density information for classifying imbalanced data," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 12, pp. 2353–2367, Dec. 2019.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[39] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.* Springer, 2005, pp. 878–887.

[40] G. Wang, T. Zhou, K.-S. Choi, and J. Lu, "A deep-ensemble-level-based interpretable Takagi– Sugeno–Kang fuzzy classifier for imbalanced data," *IEEE Trans. Cybern.*, vol. 52, no. 5, pp. 3805–3818, May 2022.

[41] A. Graves and A. Graves, "Long short-term memory," in *Supervised Sequence Labelling with Recurrent Neural Networks*, 2012, pp. 37–45.

[42] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. ACL, 2019, pp. 527–536.

[43] C. Busso et al., "LEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[44] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. ACL, 2014, pp. 1746–1751.

[45] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (volume 1: Long papers)*. ACL, 2017, pp. 873–883.

[46] M. Ren, X. Huang, W. Li, D. Song, and W. Nie, "LR-GCN: Latent relation-aware graph convolutional network for conversational emotion recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 4422–4432, 2022.

[47] T. Kim and P. Vossen, "EmoBERTa: Speaker-aware emotion recognition in conversation with roberta," 2021, *arXiv:2108.12009*.

[48] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2FNet: Multi-modal fusion network for emotion recognition in conversation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4652–4661.

[49] W. Shen, J. Chen, X. Quan, and Z. Xie, "Dialogxl: All-in-one XLNet for multi-party conversation emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 15, 2021, pp. 13789–13797.

[50] L. van der Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

[51] I. Goodfellow et al., "Generative adversarial nets," *Adv. Neural Inf. Process. Syst.*, vol. 27, no. 1, pp. 1–9, 2014.

[52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.