electronics

MDPI

Article

Multimodal Emotion Recognition Method Based on Domain Generalization and Graph Neural Networks

Jinbao Xie [1,2], Yulong Wang [2,3], Tianxin Meng [2,3], Jianqiao Tai [2,3], Yueqian Zheng [1,2,*] and Yury I. Varatnitski [4]

[1] College of Physics and Electronic Engineering, Hainan Normal University, Haikou 571158, China; jbxpost@163.com
[2] Key Laboratory of Data Science and Smart Education, Ministry of Education, Hainan Normal University, Haikou 571158, China; 202212083900012@hainnu.edu.cn (Y.W.); 202312083900014@hainnu.edu.cn (T.M.); 202313085404003@hainnu.edu.cn (J.T.)
[3] College of Information Science Technology, Hainan Normal University, Haikou 571158, China
[4] Faculty of Radiophysics and Computer Technologies, Belarusian State University, 220030 Minsk, Belarus; vorotn@bsu.by
* Correspondence: zhengyueqian@hainnu.edu.cn; Tel.: +86-18182749891

Abstract: In recent years, multimodal sentiment analysis has attracted increasing attention from researchers owing to the rapid development of human–computer interactions. Sentiment analysis is an important task for understanding dialogues. However, with the increase of multimodal data, the processing of individual modality features and the methods for multimodal feature fusion have become more significant for research. Existing methods that handle the features of each modality separately are not suitable for subsequent multimodal fusion and often fail to capture sufficient global and local information. Therefore, this study proposes a novel multimodal sentiment analysis method based on domain generalization and graph neural networks. The main characteristic of this method is that it considers the features of each modality as domains. It extracts domain-specific and cross-domain-invariant features, thereby facilitating cross-domain generalization. Generalized features are more suitable for multimodal fusion. Graph neural networks were employed to extract global and local information from the dialogue to capture the emotional changes of the speakers. Specifically, global representations were captured by modeling cross-modal interactions at the dialogue level, whereas local information was typically inferred from temporal information or the emotional changes of the speakers. The method proposed in this study outperformed existing models on the IEMOCAP, CMU-MOSEI, and MELD datasets by 0.97%, 1.09% (for seven-class classification), and 0.65% in terms of weighted F1 score, respectively. This clearly demonstrates that the domain-generalized features proposed in this study are better suited for subsequent multimodal fusion, and that the model developed here is more effective at capturing both global and local information.

Keywords: domain generalization; graph neural network; multimodal emotion recognition; domain invariant features; interdomain invariant features

Academic Editor: Arkaitz Zubiaga

Received: 12 January 2025
Revised: 10 February 2025
Accepted: 14 February 2025
Published: 23 February 2025

Citation: Xie, J.; Wang, Y.; Meng, T.; Tai, J.; Zheng, Y.; Varatnitski, Y.I. Multimodal Emotion Recognition Method Based on Domain Generalization and Graph Neural Networks. *Electronics* **2025**, *14*, 885. https://doi.org/10.3390/electronics14050885

1. Introduction

Emotions are intrinsic to humans, guiding their behavior and indicating underlying thought processes; therefore, understanding and recognizing emotions are crucial for developing artificial intelligence technologies that can interact directly with humans, such as personal digital assistants. In conversations among individuals, emotions constantly fluctuate as each person experiences and expresses them. With the rapid advancement of information technology and the popularity of social media, people generate a vast

amount of multimodal data in their daily lives, such as text, images, and audio. These data contain rich emotional information such as the emotions expressed by users in their posts, comments, pictures, and videos on social media. Consequently, automatic emotion recognition and understanding have become important research directions in natural language processing and artificial intelligence.

However, owing to the complexity and heterogeneity of multimodal data, single-modal emotion recognition methods are no longer sufficient to meet the needs of emotion analysis in conversations. Therefore, multimodal emotion recognition methods are required to better understand and analyze the emotional content of dialogues.

Furthermore, analysis of multimodal conversational data has significant applications in social life. For example, in sentiment analysis, malicious sentiment manipulation on the Internet can threaten user safety. The online sentiment environment can be effectively improved by comprehensively analyzing users' online discussions and taking appropriate measures. In intelligent customer service, systems can dynamically adjust response strategies based on the emotional content and tone expressed in user speech, thereby enhancing the user experience. In the medical industry, emotion recognition systems can assist in diagnosing and treating patients, particularly in communicating with patients with facial and body movement differences such as those with autism, providing important diagnostic and therapeutic support.

Mainstream approaches to emotion recognition in dialogue typically utilize global and local contextual information to predict the emotional labels of each sentence. Global information can be captured by modeling cross-modal interactions at the dialogue level, whereas local information is often inferred from temporal information or emotional changes of speakers. However, the extraction of single-modal features by existing methods is not conducive to subsequent multimodal fusion, resulting in limitations in emotion recognition performance. Therefore, this study proposes a novel multimodal emotion recognition method based on domain generalization and graph neural networks to improve the performance and robustness of emotion recognition.

The main contributions of this study are as follows.

(1) We proposed a multimodal sentiment recognition framework called TL-RGCN. The framework consists of three modules: the feature extraction module, where we use the RoBERTa model to extract features for the textual modality, the OpenSmile toolkit to extract features for the audio modality, and the DenseNet convolutional network to extract facial features for the visual modality. The domain-generalization module treats each modality as a separate domain and performs domain generalization across different domains. This involves splitting the features of each modality into two parts: one for extracting domain-specific invariant features and the other for domain-invariant features. The graph neural network module utilizes a Relational Graph Convolutional Network (RGCN) to capture both global and local information, thereby enhancing the performance of sentiment recognition in dialogs.

(2) We conducted experiments on three datasets: IEMOCAP, CMU-MOSEI and MELD, achieving improvements of 0.97%, 1.09% (for seven-class classification), and 0.65% in weighted F1 scores compared to existing models. (3) We performed an ablation study on the domain generalization and graph neural network modules using the IEMOCAP dataset to verify the effectiveness of both modules. When the graph neural network module was removed, the accuracy and weighted F1 score decreased by 3.69% and 3.80%, respectively. When the domain generalization module was removed, they decreased by 3.01% and 3.19%, respectively.

Overall, the proposed TL-RGCN framework and experimental results contribute to advancing the field of multimodal sentiment recognition, particularly in the context of domain generalization.

The second section of this paper focuses on related research, primarily introducing the theoretical knowledge and research status of the graph neural network module and the domain generalization module used in this study. The third section discusses the methodology, detailing the specific methods employed and the structure of the model. The fourth section covers the experiments conducted, including comparative experiments, ablation studies, and error analysis. The fifth section provides the conclusion, summarizing the methods and contributions proposed in this paper and offering insights into future research directions.

2. Related Research

This section provides a literature review of the applications of multimodal emotion recognition and graph neural networks.

2.1. Multimodal Emotion Recognition in Dialogues

The presence of multiple interlocutors, dynamic interactions, and contextual dependencies poses challenges for Emotion Recognition in Conversations (ERC) tasks. First, in a multi-participant dialogue, the emotional states of individuals can mutually influence and change each other. Relying solely on independent emotion recognition for each participant is insufficient and modeling the interactions between participants is necessary. For example, a person's emotions can be influenced by the responses of others, triggering a chain reaction. Capturing these interactive processes among participants is crucial for ERC modeling.

Second, dialogue is a dynamic interaction process, and the emotional states of participants can continually evolve as the conversations progress. Merely relying on information from the current utterance is far from sufficient; modeling the contextual information of the entire dialogue history is required. Implicit emotional attitudes, interaction patterns, and past utterances can influence current emotional states. The effective modeling of temporal dependency is also a significant challenge.

Furthermore, dialogues contain rich contextual information, including the participants' identities, relationships, topics, and environments. These contextual factors are closely related to emotions; however, comprehensive modeling is difficult. In comparison, most existing research primarily focuses on textual modalities, neglecting other potentially valuable sources of information, such as speech and visuals. Exploring how to integrate multimodal information to enhance the modeling of emotional dynamics in dialogue is a promising direction for future research.

In the ERC task, researchers strive to model the context of conversations and focus primarily on the textual modality. Notable methods include CMN [1], DialogueGCN [2], DER-GCN [3], SDT [4], DialogueCRN [5], and DAG-ERC [6].

Multimodal machine learning approaches have gained popularity because of their ability to address the limitations of unimodal methods in capturing complex real-world phenomena. The CMN [1] directly concatenates features from different modalities and utilizes Gated Recurrent Units (GRU) to model contextual information. ICON [7] extracts multimodal dialogue features and employs hierarchical modeling of emotional influences using global memory to improve the performance of speech video emotion recognition. ConGCN [8] models utterances and speakers as nodes in a graph and captures both contextual and speaker dependencies as edges. However, the ConGCN focuses only on text and audio features, neglecting other modalities. Additionally, MMGCN [9] is a model based

on Graph Convolutional Networks (GCN) that effectively captures long-range contextual information and multimodal interaction information.

Recently, Xue et al. proposed a multimodal sentiment analysis method that combines self-supervision and multilayer cross-modal attention. It utilizes multilayer cross-modal attention to enhance flexibility and information exchange between modalities [10]. Han et al. presented a multimodal sentiment classification method that first fuses image and text features to obtain multimodal sentiment classification results. The single-modal and multimodal sentiment classification results are then passed to a fully connected layer, which adjusts the dynamic weights to obtain the final sentiment classification results [11]. Lian et al. proposed a novel framework that combines semisupervised learning and multimodal interactions; however, it currently handles only two modalities, text and audio, leaving visual information for future work [12]. Shi and Huang introduced MultiEMO, which is an attention-based multimodal fusion framework that effectively integrates information from the text, audio, and visual modalities [13]. However, neither model addresses the temporal aspects of dialogue.

2.2. Graph Neural Networks

In recent years, there has been an increasing interest in representing data as graphs. However, the complexity of graph data poses challenges for traditional neural network models.

Traditional vector-based neural network models face several challenges when handling graph data. First, graph data are highly irregular and complex, making it difficult to represent them using fixed-size vectors. The numbers of nodes and edges in a graph can vary significantly, making it challenging to input them directly into neural networks. Second, graph structures contain rich topological information and neighborhood dependencies that are difficult to capture using simple vector encoding. Finally, graph data often exhibit high dynamics and heterogeneity, making them difficult to describe using a unified modeling framework.

To better leverage the structural characteristics of graph data, researchers have proposed new deep-learning models such as Graph Neural Networks (GNNs). GNNs employ convolution and pooling operations defined on graph structures to extract the features of nodes and edges in a graph and model their complex topological relationships. In addition, GNNs exhibit good scalability and flexibility, making them applicable to various types of graphical data.

Compared with traditional neural networks, GNNs demonstrate significant advantages in modeling graph data. First, GNNs can better capture the structural information of nodes and their neighborhoods in a graph, resulting in richer feature representations. Second, GNNs can leverage the topological information of a graph to model the complex interactions between nodes, which is valuable in many practical applications. Additionally, GNNs have good interpretability, clearly demonstrating the mechanisms of feature extraction and information propagation, and aiding in a deeper understanding of model behavior.

When faced with complex interdependencies between modalities, GNNs offer a more effective approach for harnessing the potential of multimodal datasets. The strength of a GNNs lies in its ability to capture and model interactions within and between modalities. This flexibility makes these attractive choices for multimodal learning tasks.

Graph Convolutional Networks (GCNs) have experienced rapid development. They optimized the node features by effectively utilizing the topological structure between nodes, improving the classification accuracy while having fewer model parameters, making it easier to train.

Extensive research has been conducted on utilizing the power of GNNs to model dialogue. DialogueGCN models dialogue using a directed graph with utterances as nodes and dependency relationships as edges, integrating them into a GCN architecture [2]. An MMGCN effectively integrates multimodal information using an undirected graph, capturing long-range context and cross-modal interactions [9]. Lian et al. proposed a GNN-based ERC architecture that utilized both text and audio modalities [12]. DialogueCRN integrates multi-turn reasoning modules to extract and integrate emotional features, allowing for a comprehensive understanding of the dialogue context from a cognitive perspective [5]. AdaIGN proposes an adaptive interactive graph network that utilizes node and edge selection strategies to balance self-dependency and empathy in conversational emotion recognition, addressing the shortcomings of existing methods in handling cross-modal emotion conflicts and contextual dependencies [14]. COGMEN employs a GNN-based architecture to model complex dependencies, including local and global information in dialogues [15]. GA2MIF introduces a multimodal fusion approach named Graph and Attention based Two-stage Multi-source Information Fusion for emotion detection in conversation. This proposed method circumvents the problem of taking heterogeneous graphs as input to the model while eliminating complex redundant connections in the construction of graphs [16]. GraphMFT proposes a graph network-based multimodal fusion technique to enhance emotion recognition in conversations, significantly improving the performance of existing models in multimodal information integration and emotion classification [17]. SDT proposes a self-distillation-based transformer model that significantly enhances multimodal emotion recognition performance by capturing intra- and inter-modal interactions in conversations and dynamically learning modality weights [4]. LLMs proposes a framework that integrates multiple emotion lexicons with advanced large language models, significantly enhancing the understanding of emotions and response generation in automated psychotherapy [18]. MFB proposes a generalized multimodal fusion method that effectively integrates Bidirectional Long Short-Term Memory networks (BiLSTM) and Bidirectional Gated Recurrent Units (BiGRU) layers for complex feature extraction [19]. UniSA proposes a unified generative framework that successfully integrates multiple subtasks of sentiment analysis through a task-specific prompt method and novel pre-training tasks, thereby enhancing the model's ability to perceive multimodal sentiment [20]. Chen et al. introduced M3Net, a multivariate, multifrequency, and multimodal graph neural network, to explore the relationships between modalities and background [21]. However, it focuses primarily on intermodality interactions and does not consider the temporal aspects within the graph.

2.3. Domain Generalization

Domain generalization is an important approach in cross-domain sentiment analysis. Sentiment analysis, as a core task in natural language processing, plays a crucial role in various practical applications such as public opinion monitoring, customer service, and product review analysis. However, owing to differences in language expressions across different domains or scenarios, sentiment analysis models based on a single domain struggle to maintain good generalization performance in cross-domain settings.

Traditional solutions often involve the fine-tuning and adaptation of models using a large amount of labeled data. However, this approach has two main problems: (1) For newly emerging domains, the collection and annotation of a substantial amount of data is often costly. (2) The models tend to overfit specific domain features, making it challenging to maintain good generalization capabilities.

Domain generalization techniques provide more efficient and robust solutions for addressing these issues. The core idea is to train a general sentiment analysis model that can be generalized to multiple domains using a limited training dataset. This requires the model

to learn shared sentiment expression features across domains without overfitting specific patterns in a particular domain. This approach improves the generalization capabilities of the model by leveraging the data relationships between the source and target domains. The existing domain generalization methods include data augmentation, domain-invariant feature learning, and meta-learning [22]. Most existing studies have focused on specific applications, such as computer vision and reinforcement learning. In this context, we primarily focus on domain-invariant feature learning methods for multimodal sentiment recognition tasks.

Domain-invariant feature learning is a popular strategy in domain generalization aimed at learning representations that remain invariant across different domains, thereby facilitating cross-domain generalization. For example, Ganin et al. [23] proposed a domain-adversarial neural network (DANN) using adversarial training, in which they attempted to confuse the domain classifier to make it unable to distinguish the domain to which the features belong, thus achieving domain-invariant learning. Similar to the DANN, many other researchers have proposed various methods to learn domain-invariant features for DG and have achieved significant success.

The popularity and effectiveness of domain-invariant learning naturally led us to explore the underlying principles behind this approach: What are the domain-invariant features in domain-invariant learning, and how can we further improve their performance in domain generalization? Previous studies by Zhao et al. [24] have shown that cross-domain feature alignment alone is insufficient to adapt to domains, and have focused on label functions. However, accessing label functions is not possible in the DG because the target is unseen. To address this issue, Bui et al. [25] utilized a metadomain-specific domain invariance.

Traditional sentiment analysis models often perform poorly in new domains because they struggle to adapt well to the data characteristics of the new domain. Therefore, designing models that can be generalized across different domains is an important research direction.

We summarized the research questions, shortcomings, performance issues, and potential improvements of the related literature, as shown in Table 1.

Table 1. Summary of Literature on Multimodal Emotion Recognition and Graph Neural Networks.

Literature	Research Problems/Disadvantages	Performance Issues	Possible Methodological Improvements
CMN [1]	Relied solely on independent emotion recognition, neglecting inter-participant emotional influences	Insufficient for capturing emotional dynamics	Developed models to capture emotional interactions among participants
DialogueGCN [2]	Primarily focused on text, neglecting multimodal data integration	Limited in capturing cross-modal interactions	Integrated audio and visual features to enhance performance
DER-GCN [3]	Did not fully consider the temporal dynamics of dialogues	Struggled with dynamic emotional state processing	Introduced temporal modeling to capture emotional evolution
SDT [4]	Lacked comprehensive context modeling over the entire dialogue history	Insufficient focus on current utterances, missing historical context	Incorporated dialogue history for a more holistic understanding
DialogueCRN [5]	Insufficient attention to emotional influences in multi-turn dialogues	Lacked the ability to understand complex emotions	Enhanced multi-turn reasoning capabilities for emotion extraction

Table 1. *Cont.*

Literature	Research Problems/Disadvantages	Performance Issues	Possible Methodological Improvements
DAG-ERC [6]	Primarily based on text, not utilizing other modalities	Missed potential insights from multimodal data	Explored multimodal integration to enrich emotional understanding
ICON [7]	Limited use of temporal dependencies	Limited understanding of dynamic interactions	Implemented temporal dependencies in multimodal feature extraction
ConGCN [8]	Focused only on text and audio, neglecting visual data	Incomplete modality coverage	Included visual features to improve emotional recognition
MMGCN [9]	Limited ability to capture long-range contextual information	Struggled with distant dependencies	Enhanced long-range context modeling to improve performance
Xue et al. [10]	Did not consider temporal aspects in sentiment analysis	Static modeling of emotional states	Introduced temporal dynamics in sentiment classification
Han et al. [11]	Poor performance in single-modal fusion	Performance dropped in complex scenarios	Explored more comprehensive multimodal fusion techniques
Lian et al. [12]	Used only text and audio, ignoring visual information	Incomplete emotional representation	Expanded to include visual data for a fuller sentiment analysis
Shi and Huang [13]	Did not address temporal dynamics in dialogues	Integration effectiveness was lacking	Developed methods to integrate temporal aspects
AdaIGN [14]	Insufficient in handling cross-modal emotional conflicts and contextual dependencies	Existing methods were inadequate for emotion recognition	Improved node and edge selection strategies to balance self-dependency and empathy
COGMEN [15]	Inadequate modeling of complex dependencies	Insufficient integration of local and global information	Enhanced modeling capabilities for complex dependencies in dialogues
GA2MIF [16]	Issues with processing heterogeneous graph inputs	Complex redundant connections affected performance	Simplified connections in graph construction and optimized information fusion methods
LLMs [18]	Limited enhancement in emotion understanding and response generation	Poor performance in complex emotional scenarios	Integrated multiple emotion lexicons with large language models to enhance emotional recognition
Chen et al. [21]	Primarily focused on intermodality interactions, neglecting temporal aspects	Lacked consideration for temporal dynamics	Focused on temporal dynamics within the graph to improve intermodality interaction modeling
Ganin et al. [23]	Existing models struggle with adaptability in new domains	Overfitting to specific domain features	Designed models that can generalize across domains to improve generalization capabilities
Zhao et al. [24]	Insufficient cross-domain feature alignment, making adaptation difficult	Inaccessibility of label functions affected model performance	Explored meta-domain-specific invariance methods
Bui et al. [25]	Traditional sentiment analysis models perform poorly in new domains	Difficulty adapting to the data characteristics of new domains	Designed sentiment analysis models that can generalize across different domains

3. Methods

To make the unimodal features more suitable for subsequent multimodal fusion and to better capture global and local information during the multimodal feature fusion process, this paper proposes a novel multimodal emotion recognition method based on domain generalization and graph neural networks, namely TL-RGCN. Figure 1 illustrates the model architecture diagram for this paper. Given a dialogue $C[u_1, u_2, \ldots, u_N]$ consisting of utterances from multiple speakers, let us represent it as sets of speakers S. Each utterance u is associated with three modalities, including text (t), audio (a), and visual (v), which can be represented as u_i^t, u_i^a, u_i^v, respectively. Using local and global context representations, the objective of the emotion recognition task in the dialogue is to predict the label $[z_1, z_2, \ldots, z_M]$ for each utterance from a set of M predefined emotion labels Z.

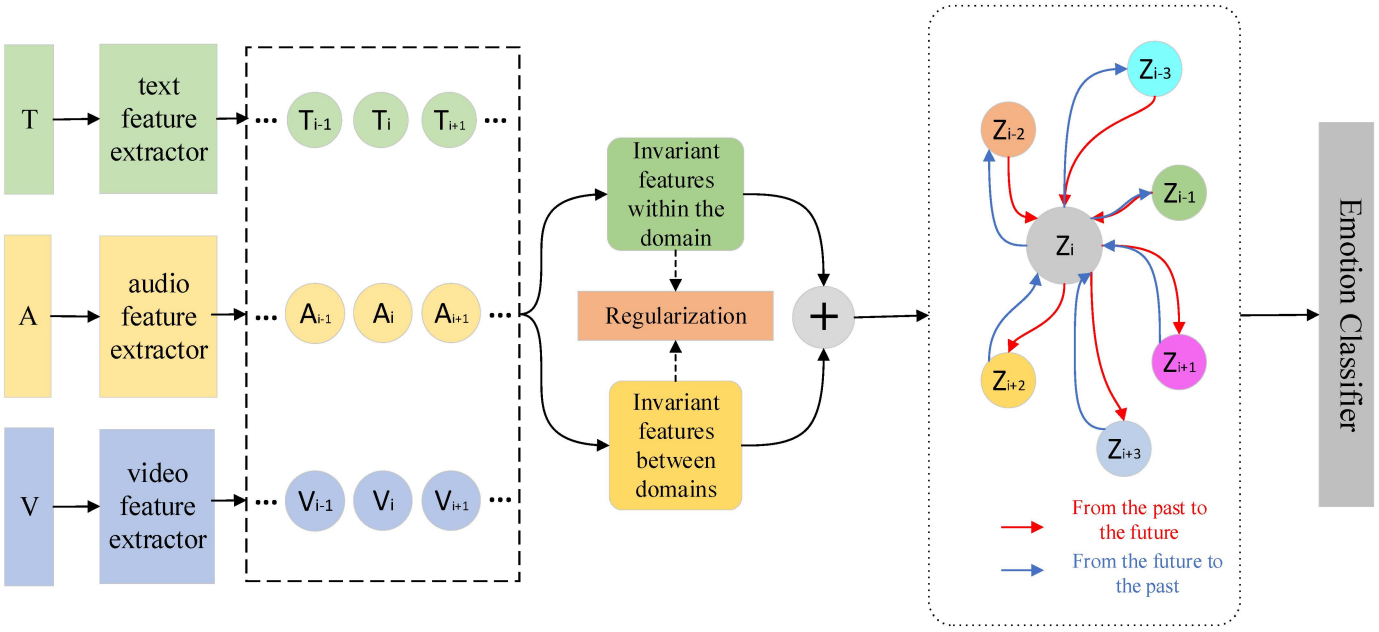

Figure 1. Model Architecture Diagram.

3.1. Feature Extraction Module

For the text modality, this study utilized the RoBERTa model for pretraining and fine-tuning to extract text features. For audio modality, the OpenSmile tool was used for feature extraction. For the video modality, DenseNet was used to extract facial expression features.

RoBERTa is an improved and optimized version of BERT (Bidirectional Encoder Representations from Transformers) model. It was pretrained on a large-scale dataset with larger training data and a longer training time, which improved the model's performance and generalization ability. The structure of the RoBERTa model is similar to that of BERT, which consists of multiple transformer blocks. Each block includes a multihead self-attention mechanism, fully connected layers, and residual connections. The multihead self-attention mechanism enables context-dependent encoding, whereas the fully connected layers map the encoded vectors to a new vector space. Specifically, the structure of the RoBERTa model is represented by Equation (1):

$$h_l = TransformerBlock(h_{l-1}) \tag{1}$$

The input vector of the RoBERTa model comprises three components: token embedding, position embedding, and segment embedding. Token embedding converts the input text into corresponding vector representations, position embedding indicates the position

information of each word in the input text, and segment embedding represents the boundary information of each sentence in the input text. Compared to the BERT model, RoBERTa is trained on a larger amount of text data, including web pages, forums, books, and news. This enables the model to capture contextual dependencies in natural languages better, thereby improving its performance.

OpenSmile is an open-source audio feature extraction toolkit that is used to extract various audio features from audio signals. The audio features extracted in this study include spectral, acoustic, and prosodic features. Mel-frequency cepstral coefficients (MFCCs) are among the most widely used spectral features, and are closely related to the auditory characteristics of human perception. To extract MFCC features, the raw audio signal must undergo preprocessing steps, including filtering, denoising, pre-emphasis, framing, and windowing. Next, the processed data are subjected to a discrete Fourier transform to obtain the spectrum. The signal spectrum is then processed using Mel filters, and the output of the Mel filters undergoes a logarithmic operation. Finally, a discrete cosine transform is applied to the logarithmic energy of the filter outputs to obtain the MFCC features. Acoustic features are typically used to measure the clarity of sound and are subjectively defined as evaluation metrics. However, because humans often exhibit phenomena such as choking, gasping, and trembling when experiencing emotional fluctuations, acoustic features are commonly used in the field of speech emotion recognition. Common acoustic features include glottal parameters, frequency and amplitude perturbations (jitter and shimmer, respectively), format frequencies, and bandwidths. Prosodic features reflect the degree of variation in speech rhythms such as stress, tempo, and duration. Therefore, prosodic features are frequently used in speech emotion analysis. Some prosodic features are related to the fundamental frequency (pitch), such as pitch frequency, mean, and variance, whereas others are related to energy, such as the short-term average energy, amplitude, and energy variation rate. Duration-related features also exist, such as short-term zero-crossing rate, speech rate, and duration.

DenseNet is a deep convolutional neural network (CNN) architecture proposed by Huang et al. in 2017 [26]. In contrast to traditional convolutional neural networks (CNNs), DenseNet adds direct connections in a densely connected manner, enabling more compact feature transfer and reuse within the network. As a densely connected neural network structure, DenseNet effectively improves the efficiency of feature propagation and parameter utilization through its unique design, thereby providing an efficient network architecture for image processing and other tasks. DenseNet was used to obtain the feature vector representations for the visual modality.

3.2. Domain Generalization Module

To maintain fairness, we divided the extracted features into two parts: one for domain-specific invariant feature learning and the other for cross-domain feature learning. In addition, to ensure feature diversity, we proposed a regularization term to maximize the differences between these two types of features.

Learning domain-specific invariant features: Domain-specific invariant features generated within each domain and unaffected by other domains primarily capture the intrinsic semantic information of the data. To obtain domain-specific invariant features, we employed a simple distillation framework for learning, as shown in Figure 2. Specifically, the teacher network uses Fourier phase information and class labels as inputs, and outputs the

Fourier phase information features for classification. The Fourier transform F(x) of single channel 2D data x is represented by Equation (2):

$$F(x)(u,v) = \sum_{h=1}^{H-1} \sum_{w=0}^{W-1} x(h,w) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)} \tag{2}$$

where u and v are the exponents and H and W represent height and width, respectively. The Fourier transform can be computed efficiently using the FFT algorithm. The phase components are represented by (3), where R(x) and I(x) denote the real and imaginary parts of F(x), respectively.

$$P(x)(u,v) = \arctan\left[\frac{I(x)(u,v)}{R(x)(u,v)}\right] \tag{3}$$

For data with multiple channels, the Fourier transform was computed separately for each channel to obtain the corresponding phase information.

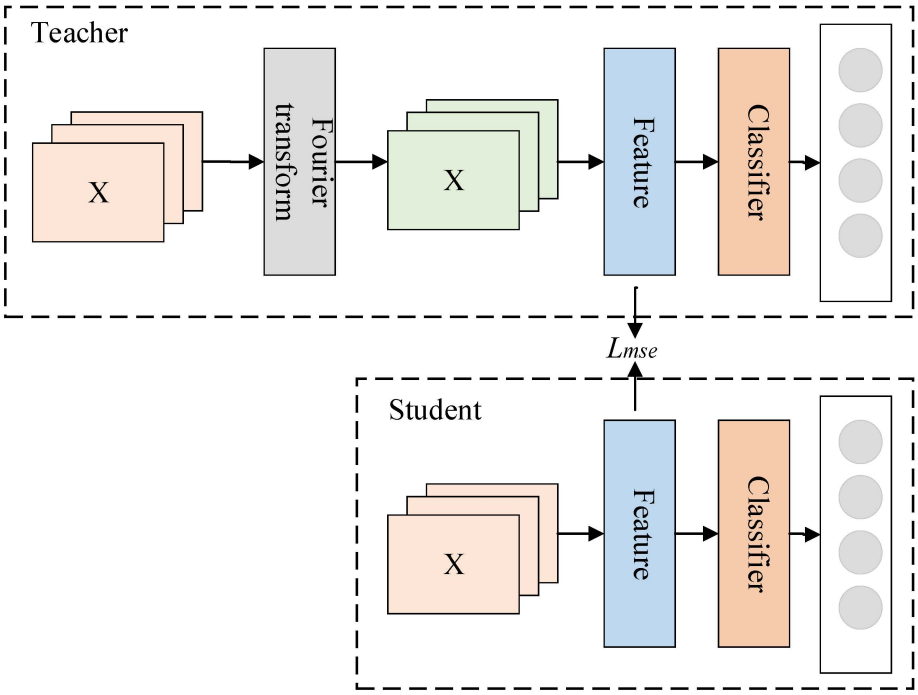

Figure 2. Teacher–Student Model Diagram.

Although the distillation method introduces additional training costs, it reduces unnecessary FFT calculations during prediction, ensuring that the entire prediction process can be performed end-to-end. The objective function of this method is given by Equation (4):

$$\min_{\theta_S^f, \theta_S^c} \mathbb{E}_{(x,y)\sim P^{tr}} \mathcal{L}_{cls}\left(G_S^c\left(G_S^f(x)\right), y\right) + \lambda_1 \mathcal{L}_{mse}\left(G_S^f(x), G_T^f(\widetilde{x})\right) \tag{4}$$

where $\mathcal{L}_{cls}$ represents the cross-entropy loss and $\mathcal{L}_{mse}$ represents the loss of the aforementioned distillation model.

Learning cross-domain invariant features: We employ the CORAL alignment to achieve this, as shown in Equation (5).

$$\mathcal{L}_{align} = \frac{2}{N \times (N-1)} \sum_{i \neq j}^{N} \left\| C^i - C^j \right\|_F^2 \tag{5}$$

To obtain diverse invariant features and reduce the redundancy and repetition between features, we aim for these two types of invariant features to be as dissimilar as possible. We achieve this using an additional regularization term.

$$\mathcal{L}_{\exp}(z_1, z_2) = -d(z_1, z_2) \tag{6}$$

Hence, the objective of this method is given as Equation (7):

$$\min_{\theta_f, \theta_c} \mathbb{E}_{(x,y)\sim P^{tr}} \mathcal{L}_{cls}\left(G_c\left(G_f(x)\right), y\right) + \lambda_1 \mathcal{L}_{mse}\left(z_1, G_T^f\left(\tilde{x}\right)\right) + \lambda_2 \mathcal{L}_{align} + \lambda_3 \mathcal{L}_{\exp}(z_1, z_2) \tag{7}$$

In a cross-domain sentiment analysis, the extraction of individual modality features is crucial. However, these individual modality features may have different distributions and representations, leading to challenges in the multimodal fusion processes. Therefore, to better adapt the extracted individual modality features to the subsequent multimodal fusion modules, we propose a graph neural network module. This module uses a graph structure to model the relationships between different modalities, enabling the adaptation and fusion of individual modality features. Specifically, we represent the features of individual modalities as nodes in a graph and utilize graph neural networks to learn the representations between nodes, thereby capturing the correlations and semantic information between modalities. Through this approach, we can better utilize the features of individual modalities and seamlessly integrate them into subsequent multimodal sentiment-analysis tasks, thereby improving the performance and generalization ability of the model.

3.3. Graph Neural Network Module

The role of the Graph Neural Network (GNN) module is to extract global and local context information and fuse the features extracted by the domain generalization module. Global information is captured by modeling cross-modal interactions at the conversation level, whereas local information is typically inferred from speaker-level temporal cues or emotional changes.

First, the domain-specific and cross-domain invariant features extracted by the domain generalization module were concatenated and input into the subsequent GNN module. Each utterance was treated as a node, and the temporal order of the utterances was used to construct a directed graph. The graph includes relationships within the same speaker's R_{intra} and relationships between different speakers' R_{inter}, as described by Equations (8) and (9).

$$R_{\text{intra}}\left(z_i^{(S_1)}\right) = \left\{ \begin{array}{l} z_i^{(S_1)} \leftarrow z_{i-P}^{(S_1)} \ldots z_i^{(S_1)} \leftarrow z_{i-1}^{(S_1)}, \\ z_i^{(S_1)} \rightarrow z_{i+1}^{(S_1)} \ldots z_i^{(S_1)} \rightarrow z_{i+F}^{(S_1)} \end{array} \right\} \tag{8}$$

$$R_{\text{inter}}\left(z_i^{(S_1)}\right) = \left\{ \begin{array}{l} z_i^{(S_1)} \leftarrow z_{i-P}^{(S_2)} \ldots z_i^{(S_1)} \leftarrow z_{i-1}^{(S_2)}, \\ z_i^{(S_1)} \rightarrow z_{i+1}^{(S_2)} \ldots z_i^{(S_1)} \rightarrow z_{i+F}^{(S_2)} \end{array} \right\} \tag{9}$$

Next, a Relational Graph Convolutional Network (GCN) module was used. The RGCN helps accumulate specific transformations of neighboring nodes based on the types and directions of the edges present in the graph using normalization and aggregation. The formula for this process is given by Equation (10).

$$x_i' = \Theta_{root} \bullet z_i + \sum_{r \in R} \sum_{j \in N_r(i)} \frac{1}{|N_r(i)|} \Theta_r \bullet z_j \tag{10}$$

where Θ_{root} and Θ_r represent the learned parameters and $|N_r(i)|$ represents the normalization constant.

The operation performed on the node feature $x'_i, x'_j (i, j \in 1, 2, \ldots n)$ obtained from the RGCN is shown in Formula (11).

$$H'_i = W_1 x'_i + \sum \alpha_{i,j} W_2 x'_j \tag{11}$$

where $\alpha_{i,j} = softmax\left(\frac{\left(W_3 x'_i\right)^T \left(W_4 x'_j\right)}{\sqrt{d}}\right)$ denotes the attention coefficient.

3.4. Multimodal Emotion Classification

Example linear layers were used to predict emotional categories, as shown in Equations (12)–(16).

$$H_i = \mathrm{Re}LU\left(W_4 H'_i + b_1\right) \tag{12}$$

$$P_i = softmax(W_5 H_i + b_2) \tag{13}$$

$$\hat{y}_i = \mathrm{argmax}(P_i) \tag{14}$$

$$\mathrm{Re}LU(x) = \begin{cases} x & if \;\; x > 0 \\ 0 & if \;\; x \leq 0 \end{cases} \tag{15}$$

$$softmax(x) = \frac{e^{x_i}}{\sum_i e^{x_i}} \tag{16}$$

To facilitate understanding, we have included some pseudocode.

```
# 1. Feature Extraction
function feature_extraction(text_data, audio_data, video_data):
    text_features = extract_text_features_using_RoBERTa(text_data)
    audio_features = extract_audio_features_using_OpenSmile(audio_data)
    video_features = extract_video_features_using_DenseNet(video_data)
    return text_features, audio_features, video_features

# 2. Learning Domain-Invariant Features
function learn_invariant_features(features):
    teacher_model = create_teacher_network()
    distilled_features = teacher_model(features)
    return distilled_features

# 3. Learning Inter-Domain Invariant Features
function learn_inter_domain_features(features):
    coral_loss = align_features_using_CORAL(features)
    return coral_loss

# 4. Feature Regularization
function apply_regularization(invariant_features1, invariant_features2):
    regularization_loss = compute_regularization(invariant_features1, invariant_features2)
    return regularization_loss

# 5. Graph Neural Network Module
function graph_neural_network(invariant_features):
    create_graph_structure(invariant_features)
    rgcn_output = process_graph_using_RGCN()
    return rgcn_output

# 6. Multimodal Emotion Classification
```

```
function multimodal_emotion_classification(graph_features):
    emotion_scores = predict_emotion_using_linear_layer(graph_features)
    return emotion_scores

# Main Program
function main(text_data, audio_data, video_data):
    text_features, audio_features, video_features = feature_extraction(text_data, audio_data, video_data)
    domain_invariant_features = learn_invariant_features(text_features + audio_features + video_features)
    inter_domain_features = learn_inter_domain_features(domain_invariant_features)
    regularization = apply_regularization(domain_invariant_features, inter_domain_features)
    graph_features = graph_neural_network(domain_invariant_features)
    emotion_scores = multimodal_emotion_classification(graph_features)
return emotion_scores

# Call Main Program
results = main(text_data, audio_data, video_data)
```

4. Experiments

This section mainly presents the relevant experiments, including the introduction of the dataset, evaluation metrics, comparative experiments (comparing with other existing models), and ablation experiments (validating the effectiveness of each module of the proposed method).

4.1. Experimental Configuration

4.1.1. Dataset Introduction

The IEMOCAP, CMU-MOSEI, and MELD datasets were used in this study. They are publicly available multimodal datasets.

IEMOCAP includes two 12 h conversation videos from 10 speakers. Each conversation is divided into words. There are 7433 utterances and 151 dialogues. The 6-channel dataset contains six emotion labels assigned to the discourse: happy, sad, neutral, angry, excited, and depressed. As a simplified version, ambiguous data pairs, such as (happy, excited) and (happy, frustrated) were merged into a 4-way dataset [27].

CMU-MOSEI is a multimodal emotion-recognition dataset. It contains annotations for 7 emotions, ranging from highly negative (-3) to highly positive ($+3$), as well as 6 emotional labels, including happiness, sadness, disgust, fear, surprise, and anger [28].

MELD is a multimodal, multispeaker dialogue dataset with three high-quality modal-aligned dialogue datasets: 13,708 utterances, 1433 dialogues, and 304 speakers. Specifically, unlike binary session datasets, such as IEMOCAP, MELD has three or more speakers in a conversation. Each sentence in the conversation was labeled with seven emotional tags: anger, disgust, fear, joy, neutrality, sadness, and surprise [29].

4.1.2. Evaluation Indicators

We used weighted F1 scores and accuracy as evaluation metrics. The F1 score is the harmonic mean of the precision and recall. If either precision or recall decreases, the F1 value decreases, and vice versa, the F1 value increases. The calculation method is shown in Equation (17).

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{17}$$

The accuracy is the percentage of correct predictions for the test set.

The weighted F1 score here refers to w-F1 in Tables 1–3, and accuracy is represented as Acc.

Table 2. Comparison of the Experimental Results on the IEMOCAP Dataset (the first six columns represent the w-F1 values for each category).

Method	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc (%)	w-F1 (%)
bc-LSTM	26.04	73.18	54.93	64.15	65.63	61.39	60.63	59.58
DialogueRNN	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75
DialogueGCN	42.75	**84.54**	63.54	64.19	63.08	**66.99**	65.25	64.18
MMGCN	45.45	77.53	61.99	66.70	72.04	64.12	64.56	64.71
DialogueCRN	51.59	74.54	62.38	67.25	73.96	59.97	65.31	65.34
COGMEN	50.79	80.73	49.24	48.83	**74.53**	62.40	65.89	65.58
GA2MIF	44.23	81.05	**65.21**	67.18	72.69	63.24	66.35	66.98
TL-RGCN	**60.76**	81.45	61.22	**68.12**	72.36	65.22	**67.90**	**67.95**

Table 3. Comparison of the Experimental Results on the CMU-MOSEI Dataset.

Method	Sentiment Classification Acc (%)		Emotion Classification w-F1 (%)					
	2 Class	7 Class	Happy	Sad	Angry	Fear	Disgust	Surprise
Multilouge-Net	81.88	43.83	66.84	64.34	66.03	86.05	73.91	85.05
TBJE	81.40	42.91	64.97	69.78	69.86	85.15	81.57	84.41
COGMEN	81.95	45.22	**69.88**	69.91	73.20	86.79	80.83	85.12
TL-RGCN	**82.66**	**46.31**	69.35	**71.86**	**75.77**	**86.90**	**83.26**	**85.48**

4.2. Comparative Experiment

This study compared models from recent years, and the results showed that the proposed model has made significant improvements. Tables 2–4 present the specific comparison results.

Table 4. Comparison of the Experimental Results on the MELD Dataset.

Method	Acc (%)	w-F1 (%)
$\text{UniSA}_{\text{GPT2}}$	48.12	31.26
bc-LSTM	59.54	56.93
DialogueRNN	59.58	57.39
DialogueGCN	59.46	58.1
DialogueCRN	60.38	58.18
GraphMFT	61.3	58.37
$\text{UniSA}_{\text{BART}}$	62.45	**60.78**
MM-DFN	62.49	59.46
SDT	62.77	60.11
TL-RGCN	**63.18**	60.76

In the IEMOCAP dataset, except for the neutral and excited categories, the proposed model outperformed the baseline model in all other categories. This means that the model had a higher accuracy and weighted F1 score in identifying emotional categories other than neutral and excited. The accuracy increased by 1.55%, indicating that the model made significant improvements in correctly classifying the samples. The weighted F1 score increased by 0.97%, indicating that the model achieved better results while balancing the accuracy and recall.

On the CMU-MOSEI dataset, except for the happy category, the model proposed in this study also outperformed the baseline model in other categories. Whether in binary classification tasks or seven classification tasks, the accuracy of this model improved by at

least 0.71% and 1.09%, respectively, compared with other comparison models. This implies that the model can more accurately classify emotional categories other than happiness, whether in binary or multiclass situations.

On the MELD dataset, the proposed model improved the accuracy and weighted F1 score by at least 0.41% and 0.65%, respectively, compared with other comparison models. This indicates that the model improved the accuracy of emotion recognition and achieved a better overall performance while balancing accuracy and recall. These results further validate the superiority of the proposed model for multimodal emotion-recognition tasks.

In summary, the proposed model demonstrated better performance than the baseline model on the IEMOCAP, CMU-MOSEI, and MELD datasets. It not only achieved better accuracy and weighted F1 scores in categories other than specific emotional categories but also demonstrated better performance in multimodal emotion recognition tasks.

In addition, we conducted a comparison with similar models DialogueGCN, MMGCN, and COGMEN, as shown in Table 5. The accuracy and w-F1 values in the table are comparisons conducted on the IEMOCAP dataset.

Table 5. Comparison with Similar Models.

Method	Time Series	Domain Generalization	Acc (%)	w-F1 (%)
DialogueGCN	×	×	65.25	64.18
MMGCN	×	×	64.56	64.71
COGMEN	×	×	65.89	65.58
TL-RGCN	√	√	**67.90**	**67.95**

4.3. Ablation Study

In this section, we describe the ablation experiments conducted to verify the effectiveness of each module.

First, we considered the influence of the modes. Table 6 presents the results obtained from the different modal combinations of the TL-RGCN model on the IEMOCAP dataset.

Table 6. Performance of the model on the IEMOCAP dataset under different modality settings.

Modality Settings	Acc (%)	w-F1 (%)
T	65.22	65.26
A	50.31	49.47
V	37.63	36.67
T + A	66.27	66.36
T + V	63.50	63.61
A + V	53.16	52.82
T + A + V	**67.90**	**67.95**

For the IEMOCAP dataset, the text modality performed the best in a single modality setting, whereas the visual modality yielded the lowest results. This may be due to the presence of noise, such as the camera position and environmental conditions. In the bimodal setting, combining text and audio modalities achieved the best performance, whereas combining audio and visual modalities yielded the worst results. The combination of the three modalities yielded the best results.

Second, the importance of each module was considered. The function of the domain generalization module is to further process the features obtained by the feature extraction module, making them better suited for subsequent feature fusion. The RGCN (Graph Convolutional Neural Network) module can better fuse the features of the three modalities and reflect relationships R_{intra} within the same speaker and R_{inter} between different speakers.

Ablation experiments were conducted to verify the effectiveness of this method. The results are summarized in Table 7.

Table 7. Performance of the model on the IEMOCAP dataset after the ablation of different modules (↓ represents the percentage point decrease compared to the original model).

Module	Acc (%)	w-F1 (%)
- $-RGCN$	64.21 (↓3.69)	64.15 (↓3.80)
- $-TL$	64.89 (↓3.01)	64.76 (↓3.19)
- $-R_{intra}$	65.54 (↓2.36)	65.82 (↓2.13)
- $-R_{inter}$	66.04 (↓1.86)	66.34 (↓1.61)
TL-RGCN	**67.90**	**67.95**

As presented in Table 7, the RGCN module has the greatest impact on the model. When it was removed, the performance of the model decreased by 3.80% (w-F1). The Domain Generalization Module (TL) is also a key part of the model and can have a significant impact on its performance, resulting in a 3.19% decrease in its w-F1 value. R_{intra} and R_{inter} are both parts of the RGCN module that complement each other and affect the performance of the model.

4.4. Explainability Analysis

We have conducted attention heatmap analysis (as shown in Figure 3) to gain deeper insights into how the model processes and fuses modalities.

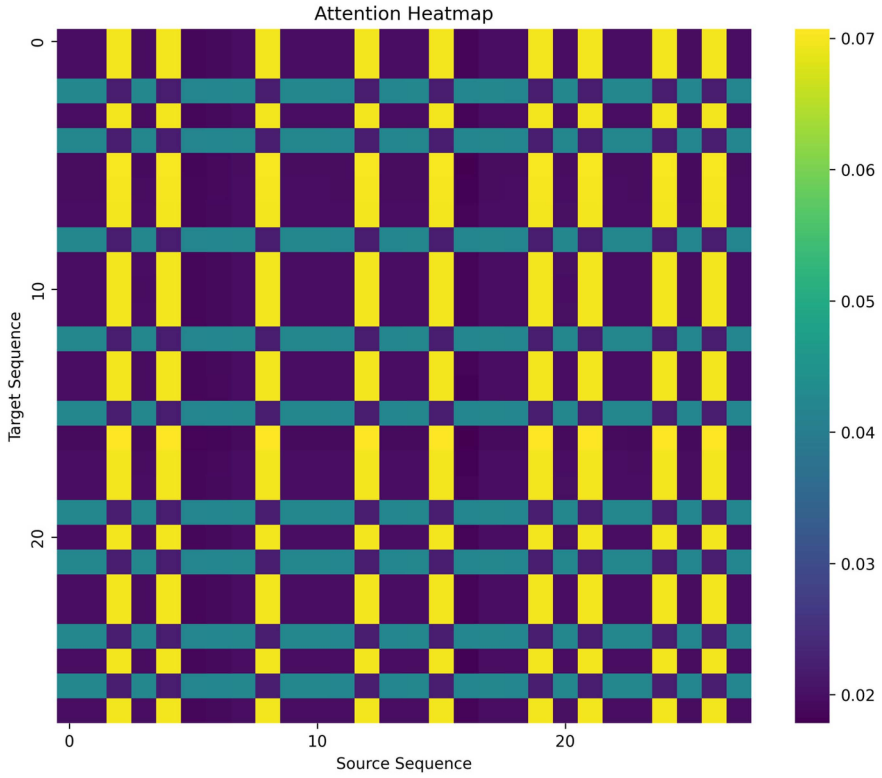

Figure 3. Confusion matrix for classification of the IEMOCAP dataset.

Through the attention heatmap analysis of the IEMOCAP dataset, we observed the following points: Firstly, the heatmap exhibits a distinct block structure, indicating strong attention between certain source sequences and target emotion categories. This structure aids in understanding how the model identifies specific emotional features. Secondly, examining the relationships between source sequences and target sequences reveals that

certain audio or visual features have a strong attention correlation with specific emotion categories. This suggests that the model effectively integrates information when fusing different modalities. For instance, when recognizing the "happy" emotion, the model may focus more on audio features relevant to that emotion. Finally, the attention distribution for some emotion categories appears more dispersed, indicating that the model may struggle with recognizing these categories. This observation provides valuable insights for future model optimization.

Additionally, we conducted a SHAP value visualization on the IEMOCAP dataset, and the results are shown in Figure 4. As shown in the figure, the audio_tensor and speaker_tensor have a significant impact on the model, while the text_tensor and visual_tensor contribute less to the overall performance.

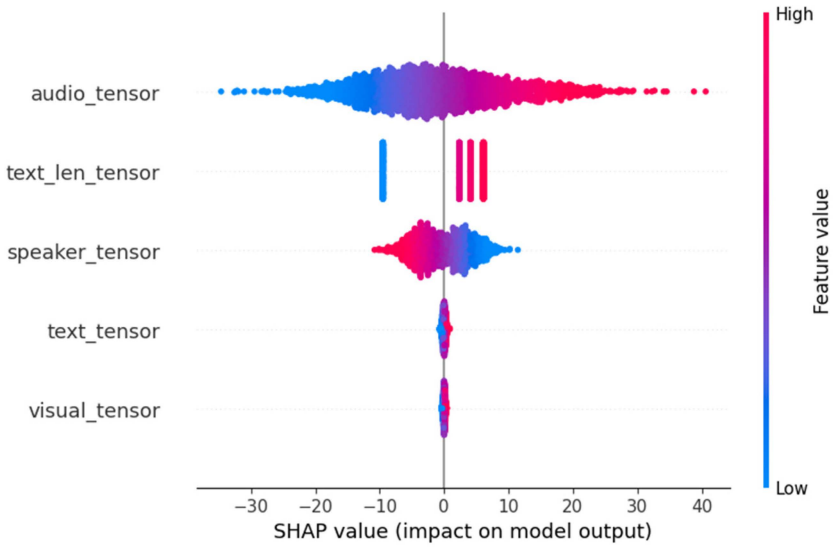

Figure 4. SHAP value visualization on the IEMOCAP dataset.

4.5. Error Analysis

This study presents a confusion matrix for the classification results of the model on the IEMOCAP dataset, as shown in Figure 5. The confusion matrix provides a comprehensive performance overview of the model for sentiment classification and allows for a detailed evaluation of the classification accuracy for each sentiment category.

After conducting predictive analysis on these datasets, we found that our model had shortcomings in distinguishing similar emotions. For example, it performed poorly in distinguishing between happiness and excitement, anger, and frustration (Figure 4). This implies that our model finds it difficult to distinguish between the subtle differences in these emotions. In addition, we observed that our model incorrectly classified other emotional labels as neutral. This may be due to the large number of samples in the neutral categories, which led our model to classify the predicted results as neutral emotions.

However, it should be pointed out that apart from the aforementioned issues, our model performs almost error-free in predicting emotions of other dissimilar categories. This indicates that our model performs relatively well at distinguishing between certain emotions.

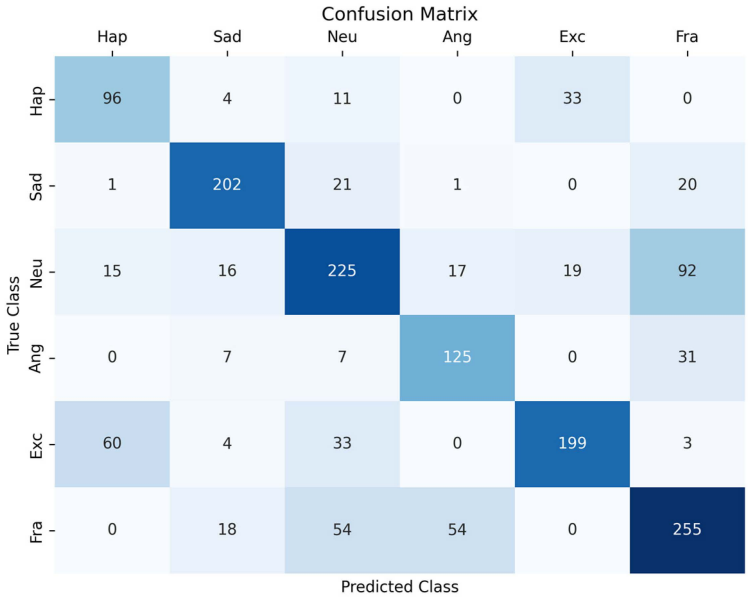

Figure 5. Confusion matrix for classification of the IEMOCAP dataset.

5. Conclusions

This paper proposes a new TL-RGCN model. The key modules included are the domain generalization and graph neural network modules. The former uses features extracted from a single modality as the domain for cross-membrane interactions, rendering the extracted features more suitable for subsequent multimodal feature fusion. The latter uses graph neural networks to learn global and local information and capture the emotions of the interlocutors. On the IEMOCAP, CMU-MOSEI, and MELD datasets, the model outperformed existing models by 0.97%, 1.09% (for seven categories), and 0.65% in weighted F1 scores, respectively. This indicates the crucial role of domain generalization in processing single-modal features, as well as the significance of the graph convolutional network module in extracting global and local information during multi-modal feature fusion. Additionally, the ablation experiments demonstrated that both the domain generalization module and the graph neural network module are indispensable; their combination significantly enhances the model's accuracy in recognizing emotional categories.

The limitations of this study include that our model exhibited some errors in distinguishing between similar categories, indicating that it struggles to differentiate the subtle differences among these emotions. Additionally, we have observed that our model erroneously classifies some other emotion labels as neutral. This may be due to the larger number of samples in the neutral category, which leads our model to be biased toward predicting neutral emotions. The future work and research directions will focus on addressing these issues by improving our model and incorporating a new module to capture the subtle differences among similar categories.

Author Contributions: Conceptualization, Y.W.; methodology, Y.W. and J.X.; validation, Y.W., J.X., T.M. and J.T.; formal analysis, Y.W. and J.X.; investigation, Y.W.; resources, J.X.; data curation, Y.W.; writing—original draft preparation, Y.W.; writing—review and editing, J.X. and Y.Z.; visualization, Y.W.; supervision, J.X., Y.Z. and Y.I.V.; project administration, J.X.; funding acquisition, J.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Education Department of Hainan Province (project number: Huky2022-19) and Key Project of Application Research on the National Smart Education Platform for Primary and Secondary Schools in Hainan Province.

Data Availability Statement: This study utilized publicly available datasets from references [27–29].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.P.; Zimmermann, R. Conversational memory network for emotion recognition in dyadic dialogue videos. In Proceedings of the Conference. Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; North American Chapter; NIH Public Access: Bethesda, MD, USA, 2018; p. 2122.
2. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv* **2019**, arXiv:1908.11540.
3. Ai, W.; Shou, Y.; Meng, T.; Li, K. DER-GCN: Dialog and Event Relation-Aware Graph Convolutional Neural Network for Multimodal Dialog Emotion Recognition. In *IEEE Transactions on Neural Networks and Learning Systems*; IEEE: Piscataway, NJ, USA, 2024.
4. Ma, H.; Wang, J.; Lin, H.; Zhang, B.; Zhang, Y.; Xu, B. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. In *IEEE Transactions on Multimedia*; IEEE: Piscataway, NJ, USA, 2023.
5. Hu, D.; Wei, L.; Huai, X. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv* **2021**, arXiv:2106.01978.
6. Shen, W.; Wu, S.; Yang, Y.; Quan, X. Directed acyclic graph network for conversational emotion recognition. *arXiv* **2021**, arXiv:2105.12907.
7. Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; Zimmermann, R. Icon: Interactive conversational memory network for multimodal emotion detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2594–2604.
8. 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 Proceedings of the IJCAI, Macao, China, 10–16 August 2019; pp. 5415–5421.
9. Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; Chua, T.S. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1437–1445.
10. Xue, K.; Xu, T.; Liao, C. Multimodal sentiment analysis network with self-supervision and multi-layer cross attention. *Comput. Appl.* **2022**, 1–6. [CrossRef]
11. Han, Y.; Ma, J. RCHFN model: An emotion classification method based on multimodal feature fusion. *Data Anal. Knowl. Discov.* **2022**, 1–16.
12. Lian, Z.; Liu, B.; Tao, J. Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition. In *IEEE Transactions on Affective Computing*; IEEE: Piscataway, NJ, USA, 2022.
13. Shi, T.; Huang, S.L. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 14752–14766.
14. Tu, G.; Xie, T.; Liang, B.; Wang, H.; Xu, R. Adaptive Graph Learning for Multimodal Conversational Emotion Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–28 February 2024; Volume 38, pp. 19089–19097.
15. Joshi, A.; Bhat, A.; Jain, A.; Singh, A.V.; 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, Seattle, WA, USA, 10–15 July 2022; pp. 4148–4164.
16. Li, J.; Wang, X.; Lv, G.; Zeng, Z. GA2MIF: Graph and attention based two-stage multi-source information fusion for conversational emotion detection. *IEEE Trans. Affect. Comput.* **2023**, *15*, 130–143. [CrossRef]
17. Li, J.; Wang, X.; Lv, G.; Zeng, Z. GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing* **2023**, *550*, 126427. [CrossRef]
18. Rasool, A.; Shahzad, M.I.; Aslam, H.; Chan, V. Emotion-Aware Response Generation Using Affect-Enriched Embeddings with LLMs. *arXiv* **2024**, arXiv:2410.01306.
19. Han, P.; Chen, H.; Rasool, A.; Jiang, Q.; Yang, M. MFB: A generalized multimodal fusion approach for bitcoin price prediction using time-lagged sentiment and indicator features. *Expert Syst. Appl.* **2025**, *261*, 125515. [CrossRef]
20. Li, Z.; Lin, T.E.; Wu, Y.; Liu, M.; Tang, F.; Zhao, M.; Li, Y. Unisa: Unified generative framework for sentiment analysis. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, USA, 29 October–3 November 2023; pp. 6132–6142.
21. Chen, F.; Shao, J.; Zhu, S.; Shen, H.T. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10761–10770.

22. Lu, W.; Wang, J.; Chen, Y.; Pan, S.J.; Hu, C.; Qin, X. Semantic-discriminative mixup for generalizable sensor-based cross-domain activity recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2022**, *6*, 1–19. [CrossRef]
23. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
24. Zhao, H.; Des Combes, R.T.; Zhang, K.; Gordon, G. On learning invariant representations for domain adaptation. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; PMLR: New York, NY, USA, 2019; pp. 7523–7532.
25. Bui, M.H.; Tran, T.; Tran, A.; Phung, D. Exploiting domain-specific features to enhance domain generalization. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 21189–21201.
26. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
27. 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. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
28. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 2236–2246.
29. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv* **2018**, arXiv:1810.02508.