Front. Digit. Educ., 2025, 2(3): 25
https://doi.org/10.1007/s44366-025-0063-x

RESEARCH ARTICLE

Emotion Dual-Space Network Based on Common and Discriminative Features for Multimodal Teacher Emotion Recognition

Ting Cai[a,b], **Shengsong Wang**[a,b], **Jing Wang**[b], **Yu Xiong**[b], **Long Liu**[b]

[a] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

[b] Research Center for Artificial Intelligence and Smart Education, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract Teacher emotion recognition (TER) has a significant impact on student engagement, classroom atmosphere, and teaching quality, which is a research hotspot in the smart education area. However, existing studies lack high-quality multimodal datasets and neglect common and discriminative features of multimodal data in emotion expression. To address these challenges, this research constructs a multimodal TER dataset suitable for real classroom teaching scenarios. TER dataset contains a total of 102 lessons and 2,170 video segments from multiple educational stages and subjects, innovatively labelled with emotional tags that characterize teacher–student interactions, such as satisfaction and questions. To explore the characteristics of multimodal data in emotion expression, this research proposes an emotion dual-space network (EDSN) that establishes an emotion commonality space construction (ECSC) module and an emotion discrimination space construction (EDSC) module. Specifically, the EDSN utilizes central moment differences to measure the similarity to assess the correlation between multiple modalities within the emotion commonality space. On this basis, the gradient reversal layer and orthogonal projection are further utilized to construct the EDSC to extract unique emotional information and remove redundant information from each modality. Experimental results demonstrate that the EDSN achieves an accuracy of 0.770 and a weighted F1 score of 0.769 on the TER dataset, outperforming other comparative models.

Received December 20, 2024; revised February 6, 2025; accepted February 21, 2025

Yu Xiong (✉)
E-mail: xiongyu@cqupt.edu.cn

Keywords teacher emotion recognition, emotion dual-space network, multimodal teacher emotion dataset, emotion commonality space construction module, emotion discrimination space construction module

1 Introduction

Teacher emotional expression in the classroom has a profound impact on teaching quality, student development and educational management, which has important educational significance and practical value (Jiang et al., 2016; Keller & Becker, 2021; Uzuntiryaki-Kondakci et al., 2022). However, in the age of AI, issues, such as technological barriers and the undervaluation of teachers' emotional competencies, are becoming increasingly prominent. Therefore, it is a key issue to accurately identify teacher emotions and promote the positive occurrence of emotional interaction, which is also the focus of this research.

Compared to the basic emotions, such as happiness and anger exhibited in daily life, teachers as a unique professional group display richer and more positive emotions during classroom instruction (Yang et al., 2024). For example, the widely used multimodal emotion datasets, such as the CMU-MOSEI (Zadeh et al., 2016; Zadeh et al., 2018), primarily encompass basic emotional expressions of speakers on various topics, often involving exaggerated and diverse reactions. These exaggerated emotion labels do not apply to relatively calm lecture scenarios. To contextualize real teacher emotions, some scholars design multimodal teacher emotion recognition (TER) datasets, such as the massive open online course

(MOOC) reviews dataset and the multimodal teacher emotion dataset (MTED) (Lu et al., 2023; Zhao et al., 2024a). Although these two datasets have made some contributions, they still have the following two limitations. First, multimodal data is somewhat default. For instance, the MOOC reviews dataset only uses two of the three modalities such as audio, text, and visual modalities, while MTED ignores visual information. Second, these two datasets neglect the emotional interaction between teachers and students. Teacher–student interaction is crucial in education, as satisfaction and questions generated during these interactions not only motivate students but also facilitate teachers' emotional expression and regulation. However, the two datasets fail to consider these unique emotional labels.

Given the above two challenges, this research constructs a multimodal TER dataset that is suitable for real classroom teaching scenarios. It covers teaching videos from multiple educational stages, including primary school, middle school, high school, and university. Moreover, it includes various subjects, such as Chinese, mathematics, English, physics, and chemistry. In total, there are 102 lessons and 2,170 video segments. To consider the specificity of teaching situations, we redefine the emotion label, which is an inheritance of and innovation from the original emotion label. On the one hand, according to the *Code of ethics for professional teachers* (Department of Education of the Republic of the Philippine, 2019), teachers should maintain a positive and healthy emotional state to guide students and try to avoid negative emotions in the teaching process (Dreer, 2024). On the other hand, some educational studies have confirmed that teacher–student interaction is an extremely important part of authentic classroom teaching, and teacher–student emotional interactions, such as satisfaction and questions, are common emotional expressions (Frenzel et al., 2021; Ong & Quek, 2023). Given these two points, this research removes the negative emotion label and then adds the teacher–student emotional interaction labels innovatively, such as satisfaction and questions. TER dataset provides rich multimodal teacher emotion data for real classroom teaching scenarios, which not only helps to optimize the quality of classroom teaching, but also effectively promotes the improvement of teachers' emotion management ability.

The analysis of the TER reveals that each modality has a different emotional polarity and contributes differently to the determination of the final teacher emotion label. Figure 1 shows the scene in which teachers ask questions with smiles. In this case, facial expression, voice intonation, and lecture text come from the same video segment and should have the same emotional polarity, such as emotion commonality (Hazarika et al., 2020). Moreover, each modality has its strengths and preferences for different emotional expressions, such as emotion discrimination (Geetha et al., 2024; Zhang et al., 2020). Specifically, the teacher's facial expression, as shown in Figure 1, is more likely to be happy, while the teacher's text directly reflects the teacher's questioning session and the voice reflects both questioning and neutral emotions. However, questioning emotions account for a greater proportion.

To simultaneously consider emotion commonality and emotion discrimination, this research proposes an emotion dual-space network (EDSN) for the multimodal TER task. It is composed of the emotion commonality space construction (ECSC) module and the emotion discrimination space construction (EDSC) module. In the ECSC, different modals' features are mapped to the emotion commonality space and the central moment discrepancy (CMD) is applied to measure the inter-modal emotion consistency to extract the common features of emotions during the multimodal fusion process (Zellinger et al., 2017). On this basis, the gradient reversal layer (GRL) and orthogonal

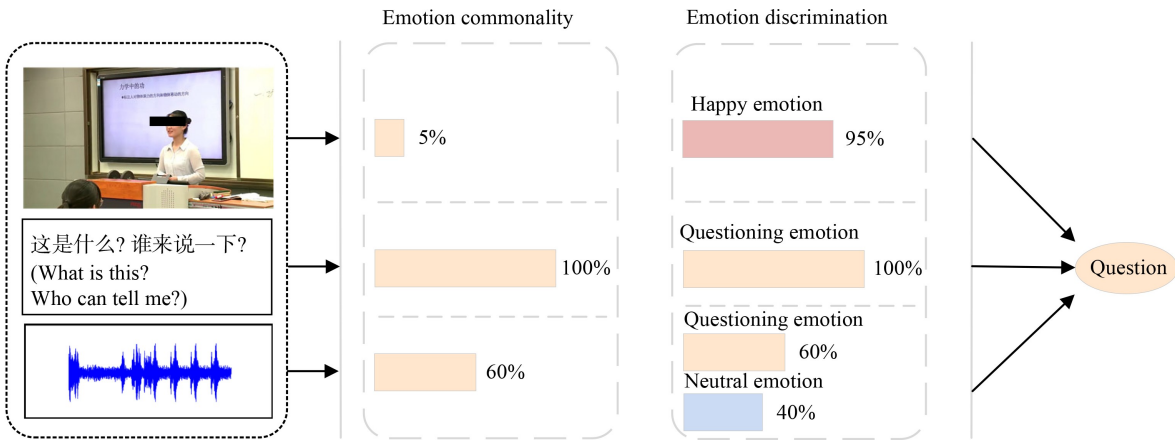

Figure 1 Multimodal emotional expression in teaching.

projection are utilized to create the inverse space of the emotion commonality space and its orthogonal space (Ganin & Lempitsky, 2015; Qin et al., 2020), which highlights the modality's emotional polarity by stripping away redundant information in the EDSC. To verify the effectiveness of the EDSN, extensive experiments are conducted on the proposed TER dataset. The experimental results show that it achieves optimal performance with an accuracy of 0.770 and a weighted F1 (WF1) score of 0.769. To further validate the generalization ability of the model, this research also conducts extensive experiments on the publicly available datasets, CMU-MOSI and IEMOCAP, which outperform the optimal model.

The main contributions of this research are summarized in the following three aspects. First, a multimodal TER dataset is constructed to adapt to real classroom scenarios. It spans multiple stages and subjects, combines multiple modalities, and proposes emotional labels that integrate teacher–student interaction innovatively. Second, the emotion commonality space is constructed to capture and extract common emotional information among different modalities by the CMD. EDSC is constructed to extract modality-specific emotional features, which highlight emotional polarity and remove redundant information through the GRL and orthogonal projection. Third, the extensive experiments conducted on the proposed TER and public datasets validate the superior performance of the proposed model over existing models.

2 Related Works

2.1 | Teacher Emotion Datasets

The construction of high-quality datasets is crucial for achieving accurate TER. The complexity of the environment, multiple subjects, and sparse expression of emotions in real classroom scenarios have led existing studies to collect only single-modal datasets (Liang et al., 2020; Wang et al., 2022). However, it is difficult for single-modal data to capture the comprehensiveness of teachers' emotions. As shown in Table 1, this research summarizes teacher emotion datasets and carries out in-depth comparative analysis from the dimensions of modal composition, labelling, quantity, and emotional interaction. For instance, Lu et al. (2023) introduced the MOOC reviews dataset, which selected different modal data based on teachers' specific behaviors, such as audio and text data for teaching as well as audio and visual modality for classroom monitoring. Zhao et al. (2024a) developed the MTED, which utilized Internet of Things cameras and smartphones in smart classrooms to record audio and text data. Although more than one data dimension is considered in the MOOC reviews dataset and MTED, there are still some modal data defaults. Teacher emotions are usually embodied by expression, intonation, and diction. Therefore, it is crucial to construct and include all the multimodalities, such as facial expressions, voice intonations, and lecture texts.

Moreover, the emotion labels annotated with teacher emotion datasets follow the basic six emotion labels, without considering the particularity of teacher–student emotional interaction. For instance, Liang et al. (2020) labelled the audio dataset with emotion labels, such as anger, sadness, happiness, neutrality, and surprise. The facial expression dataset collected by Wang et al. (2022) had additional negative emotions, such as happy, placid, restless, frustrated, angry, hateful, scared, sad, and surprised labels. The MOOC reviews dataset (Lu et al., 2023), as a multimodal dataset, only considers four labels, including happy, sad, angry, and neutral emotion labels. Basic emotion labels make it difficult to adequately capture the uniqueness of teacher emotions and the nuances of teaching situations. For example, negative emotions such as anger, hatefulness, and scare are hardly ever presented in the teaching process, while emotions such as satisfied and questioning emotion labels arising from teacher–student interactions can reflect teachers' emotions in the classroom.

Table 1 Comparison with multiple teacher emotion datasets

Dataset	Modality	Emotion label	Emotional interaction	Number
Emotion dataset (Liang et al., 2020)	a	Anger, sadness, happiness, neutrality, and surprise	×	600
Expression recognition model (Wang et al., 2022)	v	Happiness, placidness, restlessness, frustration, anger, hatefulness, scare, sadness, and surprise	×	1,250
MOOC reviews dataset (Lu et al., 2023)	(a,v) or (a,t) or (t,v)	Happiness, sadness, anger, and neutrality	×	7,622
MTED (Zhao et al., 2024a)	a, t	Happiness, surprise, neutrality, sadness, excitement, and confusion	×	5,963
TER dataset	a, v, t	Happiness, satisfaction, neutrality, and questioning	√	2,170

Notes. The abbreviations, a, v, and t, refer to audio, visual, and text, respectively. TER dataset: teacher emotion recognition dataset.

To reflect teachers' emotional changes, this research has innovatively constructed a new multimodal TER dataset. The dataset not only covers multimodal data such as facial expressions, voice intonation, and lecture texts but also takes into account emotional labels associated with teacher–student interactions, such as satisfied and questioning emotion labels.

2.2 | Multimodal Emotion Recognition

Multimodal emotion recognition aims at recognizing and understanding human emotion by integrating information from different data sources, such as audio, text, and visual sources (Khan et al., 2024). Existing multimodal emotion recognition methods primarily focus on two aspects, including multimodal feature representation and multimodal fusion methods.

The focus of feature representation is to efficiently capture features of multimodal data to improve emotion recognition. Tensor fusion network (TFN) employs outer products to learn joint representations of three modalities (Zadeh et al., 2017). Liu et al. (2018) proposed a low-rank multimodal fusion (LMF) method. LMF decomposes weights into low-rank factors, thereby reducing the number of parameters in the model. Zadeh et al. (2018) proposed the graph memory fusion network (Graph-MFN) to explore the interaction of different modalities in human multimodal language, which achieved good results while providing interpretability. Recent transformer-based contextual word representations, such as bidirectional encoder representations from transformer (BERT) and XLNet have shown impressive efficacy in natural language processing. Multimodal adaptation gate (MAG) BERT is a variant of the BERT model augmented with an MAG (Rahman et al., 2020). This allows for the integration of multimodal nonverbal data in the fine-tuning process, enabling the joint representation of visual and acoustic modalities.

Multimodal fusion networks (MFNs) are crucial in integrating diverse data sources to enhance the precision of emotion recognition. Sun et al. (2022) introduced CubeMLP, a multilayer perceptron (MLP)-based multimodal feature processing framework, which utilized three independent MLP units to combine features from three axes. Han et al. (2021) proposed the bi-bimodal fusion network, a novel end-to-end network designed to fuse and separate paired modality representations. Semantic-wise guidance for the missing modality imagination network enhances emotion recognition performance by applying modules, such as feature enrichment, adaptive fusion, and guided imagery to cope with missing modality challenges (Liu et al., 2024).

Lv et al. (2021) proposed the progressive modality reinforcement approach based on the recent advances in cross-modal transformers, which primarily addressed the issue of asynchrony between multimodal information. Yang et al. (2022b) proposed a feature-disentangled multimodal emotion recognition method to address the distribution gap and information redundancy often presented across heterogeneous modalities. Moreover, a text-oriented modality reinforcement network was proposed to mitigate potential issues stemming from variations in semantic richness among modalities (Lei et al., 2024). It emphasizes the predominant role of textual modality in multimodal language.

The existing research on multimodal emotion recognition has made some progress in the research of feature representation and multimodal fusion. However, the existing research fails to consider the emotional commonality and discrimination between different modal information. For this reason, this research proposes the EDSN. It taps into information about the common information of emotion across multiple modalities while also reinforcing the emotional bias of each modality.

3 TER Dataset

A new multimodal TER dataset is constructed to identify teachers' emotional states during the teaching process. This dataset collects videos of real classroom instruction from multiple MOOC platforms across multiple teaching stages and disciplines. The whole process consists of two phases, including data collection and emotion annotation.

3.1 | Data Collection

TER dataset is derived from multiple publicly available MOOC platforms and follows the terms of service and legal requirements of their respective platforms. To protect participants' identities, all videos are anonymized to remove personally identifiable information and sensitive contextual elements to ensure legal and ethical compliance. This research follows random sampling to balance gender, age, and subject to further ensure the generalization and fairness of the TER dataset.

After collecting the raw video, this research follows certain rules to select the segments that reflect the teachers' emotional changes and remove segments that contain redundant and distracting information. The selection of video segments and subsequent annotation is done on the proposed self-developed system (Cai et al., 2024). Three rules for segment selection are as follows. First, to ensure that each

segment accurately reflects teachers' emotions, each video segment should be concise, thereby the length should be maintained between three and ten seconds. Second, to avoid confusion among multiple targets in the video segment, each segment should only display the face and voice of the teacher, the research removes segments that contain students' faces and voices. Third, to ensure the uniqueness of teachers' emotions, video segments with emotion labels that could not be easily determined are eliminated, and segments with a distinct and single emotion polarity are retained.

Finally, this research selects 2,170 video segments from 102 open courses. Each video is available in 720p with higher definition resolution 1,280 pixels × 720 pixels, and the audio of the instructor's lectures is clear and easy to understand. The number of raw video and video segments across different educational stages is shown in Figure 2. In primary, middle, and high school, approximately 30 videos are collected from each stage, resulting in around 600 selected segments for each stage. The number of videos in universities is relatively small compared to other educational stages. This is because university instructors pay more attention to explaining knowledge points and less to mobilizing the classroom atmosphere and increasing interaction links. There are minimal fluctuations and instructors generally exhibit neutral emotions among the teachers. Therefore, fewer videos are collected from the university level in the TER dataset.

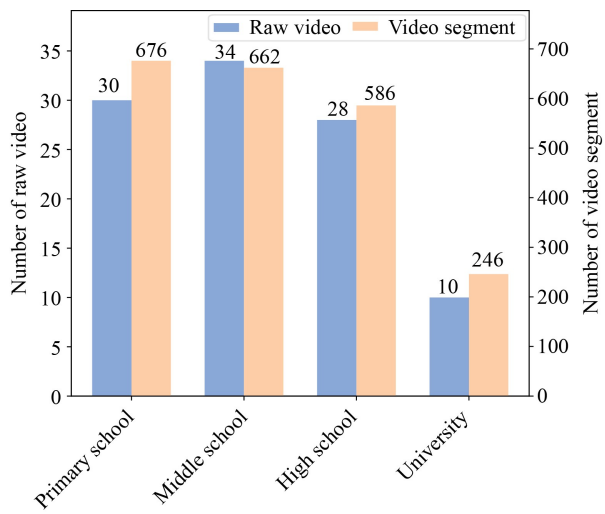

Figure 2 Numbers of raw video and video segment across different educational stages.

3.2 | Emotion Annotation

Teachers' emotional expression is different from the basic emotional labels in daily life. For example, Ekman (1984) proposed six basic emotion labels, namely, happiness, surprise, sadness, hatefulness, anger, and scare emotions, which might not be fully applicable to classroom teaching scenarios. Specifically, according to the code of ethics for teachers (Dreer, 2024), teachers need to maintain a healthy emotional state to influence students with a good state of mind and create a positive classroom atmosphere. This requires teachers to avoid negative emotions in the teaching process. Similarly, the frequency of negative emotions, such as sadness and anger, is extremely low, as confirmed in the collection of teaching videos. Therefore, this research removes negative emotion labels, such as anger and sadness. More importantly, some educational studies confirm that teacher–student interaction is an important part of real classroom teaching scenarios (Frenzel et al., 2021; Ong & Quek, 2023). In this process, teachers have a higher response frequency to students' interactive emotions such as questioning and satisfied emotion labels. For this reason, this research adds two new labels for teacher–student emotion interactions, questioning and satisfied labels. Given these two points, this research optimizes and innovates the traditional teachers' emotion labels, and ultimately determines teachers' emotion labels, including happy, satisfied, neutral, and questioning labels.

To achieve rapid and accurate emotion annotation, this research has independently developed a multimodal emotion annotation system (Cai et al., 2024). Initially, complete course videos are imported into the annotation system. The start and end times of each target video segment are determined based on the cropping rules. On this basis, three experts are invited to label the teachers' emotions for each video segment and conduct emotion recognition training to ensure a scientific and reasonable labelling process. If the annotation results from the three experts agree, the outcome is considered final without question. In the case where two annotators agree but the third disagrees, the minority follows the majority. If all three annotations disagree, arbitration from experts is sought in the emotion recognition area. Eventually, the selected segments are edited according to their start time and end time to obtain video segments with accurate emotional labels.

The specific number distributions of the teachers' emotion labels are shown in Table 2, where the data distributions for the four emotion labels are relatively balanced, with each label containing between 450 and 700 segments. On this basis, three different modal examples for each label are further given in Figure 3. It can be seen that the different modalities of the same video segment share both common emotional polarity and emotional specificity. Therefore, this research needs to consider both the emotion commonality and emotional discrimination features among multimodalities to improve the accuracy of TER dataset.

Table 2 TER dataset emotion categories

Emotion label	Number of labels	Description	Example
Happiness	474	Teachers feel happy in the teaching process.	Figure 3(a)
Satisfaction	455	Teachers feel satisfied when students perform well and learn effectively.	Figure 3(b)
Neutrality	581	Teachers exhibit calmness, showing neither positive nor negative emotions.	Figure 3(c)
Questioning	660	Teachers' emotions guide students to think and answer.	Figure 3(d)

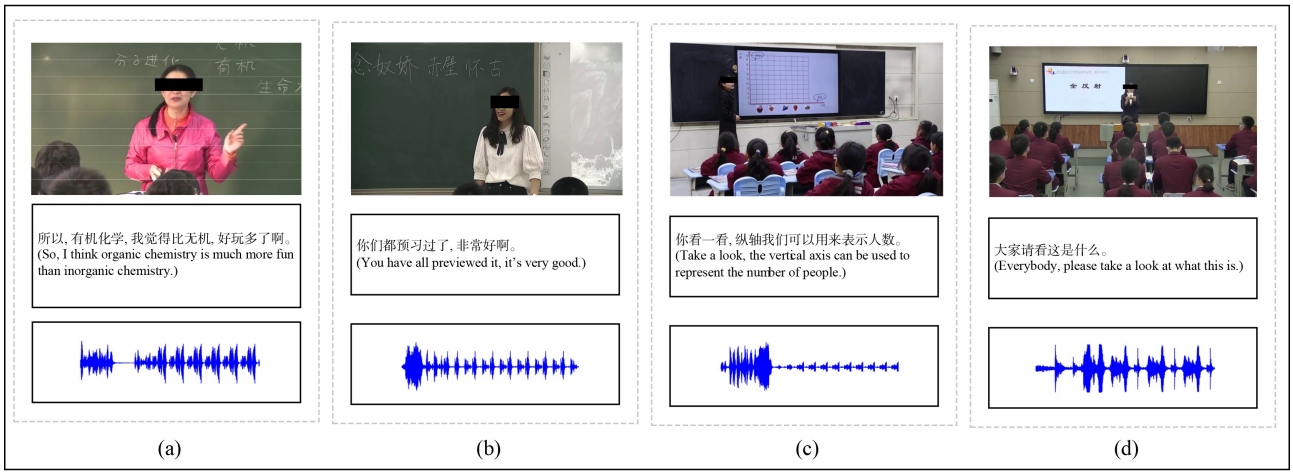

Figure 3 Some examples of TER. (a) Happiness, (b) satisfaction, (c) neutrality, (d) questioning.

4 Methodology

The overall framework of the EDSN is shown in Figure 4, including a multimodal feature extraction module, an ECSC module and an EDSC module. First, teachers' facial expressions, voice intonations, and lecture texts are processed to extract the feature representations of each modality. Second, ECSC provides an emotion commonality space, which employs the CMD to extract common information between different modalities. Third, EDSC provides an emotion discrimination space and utilizes a GRL to create an inverse space of the emotion commonality space and an orthogonal space, which helps remove redundant features and extract emotion discrimination features.

4.1 | Multimodal Feature Extraction Module

Each teaching video segment is composed of three modalities including audio ($\boldsymbol{S}_a$) through intonations and prosody, text ($\boldsymbol{S}_t$) in the form of lecture texts, and visual ($\boldsymbol{S}_a$) via facial expression, here $\boldsymbol{S}$ indicating the set of collection. To ensure a consistent sample size across three modalities, this research utilizes temporal alignment, where visual, audio, and text data from the same period are used as the same multimodal data sample. However, since each modality has its unique

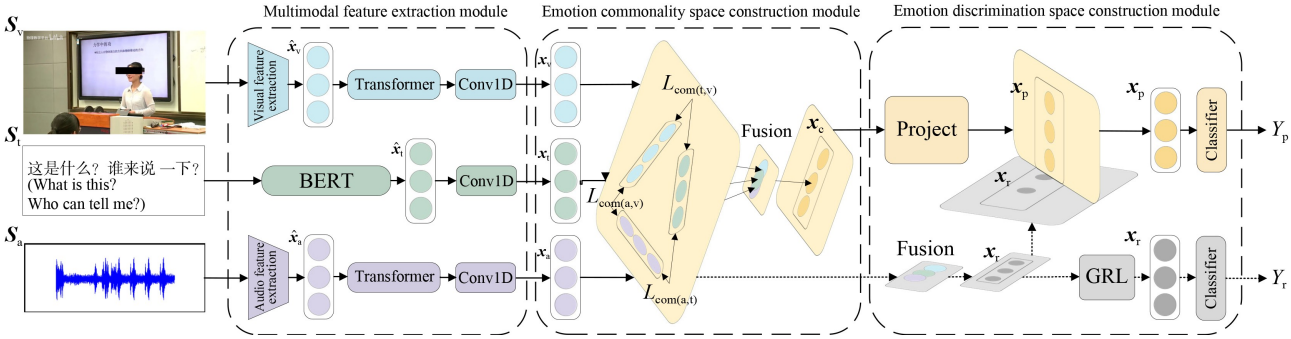

Figure 4 Overall framework of emotion dual-spatial network. BERT: bidirectional encoder representation from transformers; GRL: gradient reversal layer.

characteristics, it is necessary to use different emotion feature extraction methods.

4.1.1 Visual Feature Extraction

Facial expression is the most intuitive expression of emotion in the video segments. The process of extracting facial expression features includes three main stages: First, the number of duplicated frames in the video segments is removed by downsampling, and then the dlib face detector is used to detect faces in each frame of the video segments to obtain 68 face landmarks; second, the coordinates of face landmarks are converted into an array format to obtain 136-dimensional sequence features; third, the facial key landmark features $\widehat{\boldsymbol{x}}_{\mathrm{v}}$ of multiple image data are merged and fed into the transformer where $\boldsymbol{x}$ indicates the facial feature variant and v means the land-mark feature variant (Vaswani et al., 2017). This is aimed at exploring the intricate dependencies among various facial features and the corresponding output vectors are represented as the final visual features. The specific formula is expressed as:

$$\boldsymbol{x}_{\mathrm{v}} = \mathrm{Conv1D}(\mathrm{Transformer}(\widehat{\boldsymbol{x}}_{\mathrm{v}})), \tag{1}$$

where $\widehat{\boldsymbol{x}}_{\mathrm{v}} \in \mathbb{R}^{T_{\mathrm{v}} \times d_{\mathrm{v}}}$ represents visual features, T_{v} and d_{v} are the length and dimension of the visual sequence, respectively, here $d_{\mathrm{v}} = 136$; $\boldsymbol{x}_{\mathrm{v}} \in \mathbb{R}^{d_m \times 1}$ represents the final visual feature, d_m represents the feature dimension where m means multimodal feature. In Eq. (1), the role of the 1-dimensional convolutional layer is to perform dimensional transformation.

4.1.2 Audio Feature Extraction

Referring to the audio feature extraction in Chinese single-and multi-modal sentiment (Yu et al., 2020), librosa audio toolkit is utilized to extract audio features at a sampling rate of 22,050 Hz (McFee et al., 2015). The acoustic features comprise a 33-dimensional feature vector, which includes a 1-dimensional logarithmic fundamental frequency, 20-dimensional Mel-frequency cepstral coefficients, and 12-dimensional constant-Q transform chroma features. Finally, $\widehat{\boldsymbol{x}}_{\mathrm{a}}$ is input into the transformer to explore the relationships within the audio sequence, thereby obtaining $\boldsymbol{x}_{\mathrm{a}}$,

$$\boldsymbol{x}_{\mathrm{a}} = \mathrm{Conv1D}(\mathrm{Transformer}(\widehat{\boldsymbol{x}}_{\mathrm{a}})), \tag{2}$$

where $\widehat{\boldsymbol{x}}_{\mathrm{a}} \in \mathbb{R}^{T_{\mathrm{a}} \times d_{\mathrm{a}}}$ represents an audio feature, T_{a} (the duration of each audio) and d_{a} (the exact dimension of each audio) are the length and dimension of each audio, respectively, here $d_{\mathrm{a}} = 33$; $\boldsymbol{x}_{\mathrm{a}} \in \mathbb{R}^{d_m \times 1}$ represents the final audio features.

4.1.3 Text Feature Extraction

All audio data is transcribed by Tencent Cloud's automatic speech recognition and corrected to obtain textual data manually. Subsequently, word segmentation is performed by the Jieba. Then, we use the pretrained BERT-base-chinese to get word vectors from these segmented textual data $\boldsymbol{S}_{\mathrm{t}}$. The following is the corresponding formula expression:

$$\widehat{\boldsymbol{x}}_{\mathrm{t}} = \mathrm{BERT}(\boldsymbol{S}_{\mathrm{t}}), \tag{3}$$

$$\boldsymbol{x}_{\mathrm{t}} = \mathrm{Conv1D}(\widehat{\boldsymbol{x}}_{\mathrm{t}}), \tag{4}$$

where $\widehat{\boldsymbol{x}}_{\mathrm{t}} \in \mathbb{R}^{d_{\mathrm{t}} \times 1}$ represents the word vector obtained from the BERT model, and d_{t} represents the word vector dimension, here $d_{\mathrm{t}} = 768$; $\boldsymbol{x}_{\mathrm{t}} \in \mathbb{R}^{d_{\mathrm{t}} \times 1}$ represents the final text feature.

4.2 | Emotion Commonality Space Construction Module

Teachers' facial expression, voice intonation, and lecture text in the same video segments should present a common emotional polarity, even though they are different expressions of the same emotion. For this reason, an emotion commonality space is constructed to extract common features of emotional expression in different forms.

Specifically, CMD is utilized to measure the similarity between different modalities to assess the emotional correlation among them (Zellinger et al., 2017). CMD evaluates the distribution similarity of two vectors by calculating the similarity between the central moments of different orders. This research combines three modalities in the video in pairs and calculate the common features between each pair of modalities. For example, two different modalities of feature vector $\boldsymbol{X}$ and $\boldsymbol{Y}$ defined over a closed interval $[\boldsymbol{a}, \boldsymbol{b}]$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent visual features and textual features, respectively. $\mathbb{E}(\boldsymbol{X})$ and $\mathbb{E}(\boldsymbol{Y})$ denote the expectations of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The formula for the central moment can be expressed as follows:

$$\boldsymbol{c}_k(\boldsymbol{X}) = \mathbb{E}((\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))^k). \tag{5}$$

The central moment reflects the characteristics of data distribution, such as offset, dispersion, symmetry, and kurtosis. The similarity of the $\boldsymbol{X}$ and $\boldsymbol{Y}$ distributions is described by comparing the differences in the central moments of $\boldsymbol{X}$ and $\boldsymbol{Y}$ at different orders. In this research, the fourth-order central moment ($k = 4$) and the calculation process of CMD is as follows:

$$\mathrm{CMD}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{|\boldsymbol{b} - \boldsymbol{a}|} \|\mathbb{E}(\boldsymbol{X}) - \mathbb{E}(\boldsymbol{Y})\|_2 + \sum_{k=2}^{k} \frac{1}{|\boldsymbol{b} - \boldsymbol{a}|^k} \|\boldsymbol{c}_k(\boldsymbol{X}) - \boldsymbol{c}_k(\boldsymbol{Y})\|_2. \tag{6}$$

In this research, the CMD loss between each

pair of modalities:

$$L_{\text{com}} = \frac{1}{3}(\text{CMD}(\boldsymbol{x}_{\text{a}}, \boldsymbol{x}_{\text{t}}) + \text{CMD}(\boldsymbol{x}_{\text{t}}, \boldsymbol{x}_{\text{v}}) + \text{CMD}(\boldsymbol{x}_{\text{a}}, \boldsymbol{x}_{\text{v}})). \tag{7}$$

As the training progresses, the CMD distance of multiple modal features in the emotion commonality space gradually decreases, indicating that the emotional features of each modality tend to be consistent. Finally, the emotion commonality feature $(\boldsymbol{x}_{\text{c}} \in \mathbb{R}^{d_m \times 1})$ is represented as:

$$\boldsymbol{x}_{\text{c}} = \boldsymbol{w}(\boldsymbol{x}_{\text{a}} + \boldsymbol{x}_{\text{t}} + \boldsymbol{x}_{\text{v}}) + \boldsymbol{b}, \tag{8}$$

where $\boldsymbol{w}$ and $\boldsymbol{b}$ represent the weights and bias, respectively.

4.3 | Emotion Discrimination Space Construction Module

In addition to the emotional features, there are significant differences in emotional expression abilities across different modalities. Moreover, each modality has special emotional polarity and redundant information. For example, the happiness label is mainly reflected through facial expressions, while the questioning emotion label is more reflected in the rhythm and semantic audio expressions. Therefore, how to remove the emotional redundancy in modality and extract the emotion discrimination features in different modalities is the key to improving teachers' emotional recognition.

For this purpose, the research constructs the inverse emotion commonality space. Initially, an auxiliary fusion module is added to extend information, with the same structure as the previous fusion module but with different parameters. This allows to obtain a parallel space of emotion commonality and new fusion results $\boldsymbol{x}_{\text{r}} \in \mathbb{R}^{d_m \times 1}$. Subsequently, GRL is integrated to extract the emotional redundancy information $\boldsymbol{x}_{\text{r}}$. As illustrated in Figure 5, GRL does not operate for forward propagation. However, during the backward propagation, it reverses the sign of the gradient. This encourages $\boldsymbol{x}_{\text{r}}$ to carry more redundant information after training.

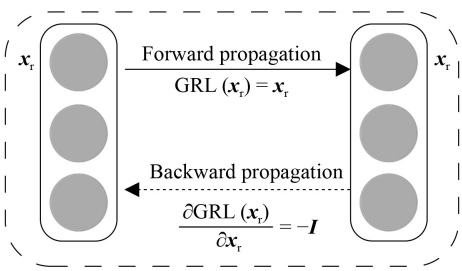

Figure 5 Working principle of the GRL. GRL: gradient reversal layer. $\boldsymbol{I}$ means the unit mareix.

After passing through the GRL, the features $\boldsymbol{x}_{\text{r}}$ are input to the classification layer. Finally, the cross-entropy loss function (L_{r}) is utilized to measure the discrepancy between the predicted and the true emotion labels $\boldsymbol{Y}_{\text{truth}}$ where $\boldsymbol{w}_{\text{r}} \in \mathbb{R}^{4 \times d_m}$ and $\boldsymbol{b}_{\text{r}} \in \mathbb{R}^{4 \times 1}$ represent weight and bias respectively:

$$L_{\text{r}} = \text{CrossEntropy}(\boldsymbol{Y}_{\text{truth}}, \boldsymbol{w}_{\text{r}}\boldsymbol{x}_{\text{r}} + \boldsymbol{b}_{\text{r}}). \tag{9}$$

On this basis, the orthogonal space of the inverse space is constructed as the emotion discrimination space to extract the unique emotion features in different modalities. Specifically, the redundant feature $\boldsymbol{x}_{\text{r}}$ does not contain valid emotional information. The feature space orthogonal to it should comprise features that are purely discriminative for classification purposes. Hence, projecting $\boldsymbol{x}_{\text{c}}$ onto the orthogonal space of $\boldsymbol{x}_{\text{r}}$, preserves the original information applicable to emotion recognition, but also removes redundant features that are irrelevant or harmful to emotion.

As shown in Figure 6, the concept of orthogonal projection is illustrated by visualizing it in 3D space. $\boldsymbol{x}_{\text{c}}$ represents the common feature derived through the ECSC, while $\boldsymbol{x}_{\text{r}}$ denotes the redundant feature obtained through the inverse space. The specific formula is as follows:

$$\boldsymbol{x}_{\text{r}}^{\text{norm}} = \frac{\boldsymbol{x}_{\text{r}}}{|\boldsymbol{x}_{\text{r}}|_2}, \tag{10}$$

$$\boldsymbol{x}_{\text{r}}^{*} = \boldsymbol{x}_{\text{r}}^{\text{norm}} \cdot (\boldsymbol{x}_{\text{c}} \cdot \boldsymbol{x}_{\text{r}}^{\text{norm}}), \tag{11}$$

$$\boldsymbol{x}_{\text{p}} = \boldsymbol{x}_{\text{c}} - \boldsymbol{x}_{\text{r}}^{*}, \tag{12}$$

where $\boldsymbol{x}_{\text{r}}^{\text{norm}}$ is the unit feature after normalizing $\boldsymbol{x}_{\text{r}}$, $\boldsymbol{x}_{\text{r}}^{*}$ is the projection feature of $\boldsymbol{x}_{\text{c}}$ onto the $\boldsymbol{x}_{\text{r}}^{\text{norm}}$, $\boldsymbol{x}_{\text{p}}$ is the projection feature of $\boldsymbol{x}_{\text{c}}$ orthogonal to $\boldsymbol{x}_{\text{r}}^{*}$. Finally, the research uses $\boldsymbol{x}_{\text{p}}$ for the final emotion classification:

$$L_{\text{p}} = \text{CrossEntropy}(\boldsymbol{Y}_{\text{truth}}, \boldsymbol{w}_{\text{p}}\boldsymbol{x}_{\text{p}} + \boldsymbol{b}_{\text{p}}). \tag{13}$$

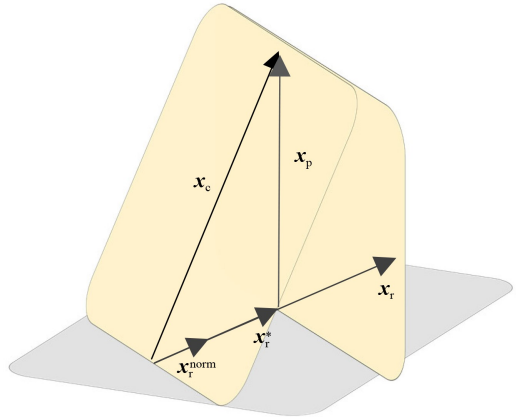

Figure 6 Mechanism of the orthogonal projection.

During the training process, the gradients of the two loss functions L_r and L_p are propagated through $\boldsymbol{x}_r$ and jointly influence it. By balancing the influence of these two factors, $\boldsymbol{x}_r$ becomes more aligned with the true redundant features. Moreover, $\boldsymbol{x}_p$ becomes more discriminative in the direction orthogonal to $\boldsymbol{x}_r$.

Ultimately, the overall loss (L) of EDSN is as follows:

$$L = L_{com} + L_r + L_p. \tag{14}$$

5 Experiments

5.1 | Experimental Setup

5.1.1 Baselines

To validate the effectiveness of the proposed model, five models, including TFN, LMF, Graph-MFN, decoupled multimodal distillation (DMD), and multimodal information bottlenecks (MIB), are compared with other state-of-the-art multimodal models. TFN utilizes outer product to learn the joint representation of three modalities, thereby enhancing its capacity to handle multi-modal data (Zadeh et al., 2017). LMF reduces high-order tensors into low-rank factors for efficient fusion (Liu et al., 2018). Graph-MFN replaces the original fusion component in the memory fusion network with the new dynamic fusion graph to achieve better interpretability and efficiency (Zadeh et al., 2018). DMD mitigates the issue of inherent multimodal heterogeneities and varies modality contributions by enabling flexible and adaptive cross-modal knowledge distillation to enhance the discriminative features of each modality (Li et al., 2023). MIB inherits from the general information bottleneck and aims to learn minimal sufficient representations for a given task by maximizing multimodal information between representations and targets (Mai et al., 2022), while constraining mutual information between representations and input data.

5.1.2 Implementation Details

To ensure the fairness of the experiments, all models are established on the PyTorch toolbox and trained on a GeForce RTX 4090 general processing unit with 24 GB memory. The Adam optimizer is employed with a learning rate of 10^{-5}, sets the batch size to 64, and trains the models for 40 epochs. The TER is divided into training, validation, and testing sets with a ratio of 7:1.5:1.5, ensuring class balance through stratified sampling. Each experiment is repeated 5 times, and the average results are reported. The best model is saved based on validation performance and loaded for testing.

5.1.3 Evaluation Metrics

For the classification task conducted on the TER dataset, three evaluation metrics are utilized: First, accuracy measures the proportion of correctly classified instances; second, F1 score, the harmonic average of precision and recall, combines the precision and completeness of the model; third, WF1 score gives different weights to each category based on the number of samples in each category and is often used in datasets where the distribution of categories is not balanced.

5.2 | Overall Comparison

As shown in Table 3, EDSN outperforms existing methods in all four emotional categories and overall recognition performance. Despite the high baseline achieved by MIB, the proposed EDSN remains state-of-the-art in terms of accuracy (ACC) and WF1 score. Specifically, EDSN achieves an ACC of 0.770 and WF1 score of 0.769, which are 5.2% and 5.3% higher than the optimal MIB, respectively. This indicates that EDSN significantly enhances the effectiveness of multimodal TER by extracting both emotion common and discriminative features across different modalities.

As can be seen in Table 3, happy and satisfied emotion labels are identified with slightly lower accuracy than neutral and questioning emotion labels.

Table 3 Comparison among recognition models

Model	Happiness		Satisfaction		Neutrality		Questioning		Total emotions	
	ACC	F1 score	ACC	F1 score	ACC	F1 score	ACC	F1 score	ACC	WF1 score
TFN	0.631	0.602	0.452	0.559	0.716	0.660	0.817	0.792	0.669	0.664
LMF	0.631	0.590	0.546	0.630	0.716	0.673	0.785	0.793	0.681	0.681
Graph-MFN	0.585	0.603	0.603	0.642	0.695	0.677	0.806	0.773	0.684	0.682
DMD	0.631	0.645	0.575	0.637	0.758	0.746	0.796	0.740	0.702	0.700
MIB	0.569	0.617	0.644	0.676	0.779	0.725	0.817	0.804	0.718	0.716
EDSN	**0.701**	**0.719**	**0.685**	**0.730**	**0.800**	**0.768**	**0.849**	**0.836**	**0.770**	**0.769**

Note. The best results are highlighted in bold.

To further analyze the reasons for this, this research plots the confusion matrices for MIB and EDSN, as shown in Figure 7. It is observed that there are two main reasons for the slightly lower recognition accuracy of happy and satisfied emotion labels. First, although they have different educational significance, they are both relatively positive emotions with similar emotional characteristics, which are easy to confuse and lead to misclassification. Second, teachers usually maintain a relatively calm state when they express happy and satisfied emotions during the teaching process, instead of more exaggerated emotions, resulting in a lower level of discrimination from the neutral emotions are less distinguishable. As a result, happiness and satisfaction of teachers are more likely to be misidentified, while neutrality and question are identified with relatively high accuracy.

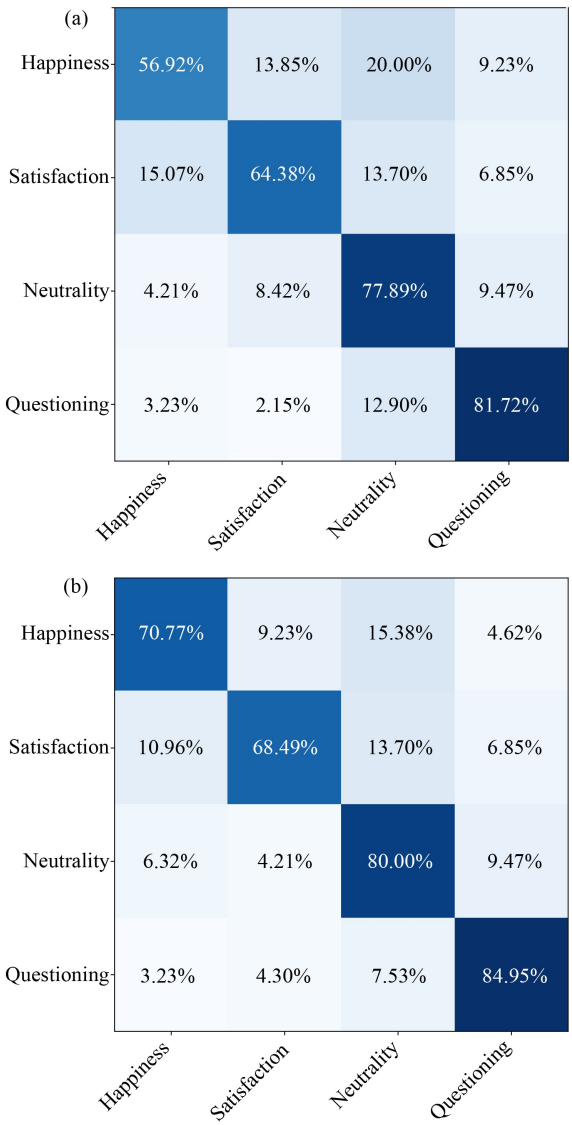

Figure 7 Confusion matrix of (a) MIB and (b) EDSN on TER dataset.

To cope with that limitation, the direction of improvement should be further considered. First, the characteristics of happiness and satisfaction are analyzed. Happiness is a subjective and spontaneous emotion for teachers, while satisfaction depends more on students' feedback. Therefore, students' feedback is added to further differentiate between happy and satisfied emotions. Second, the algorithm continues to be optimized to amplify the emotion changes powerfully, thus enhancing the model's ability to capture low-intensity emotional cues.

5.3 | Ablation Study

5.3.1 Effectiveness of ECSC and EDSC

ECSC and EDSC are two very important modules in EDSN, which are used to extract emotion commonality features and emotion discrimination features among multimodalities respectively. To verify the effectiveness of these two modules, the ablation experiments are conducted to remove the ECSC and EDSC modules from the model and analyze their respective contributions. The experimental results are shown in Table 4.

First, the EDSC model retains the EDSC module and removes the ECSC module, and then extracts only the emotion common features, resulting in a 2.5% and 2.5% decrease in overall ACC and WF1 scores, respectively. Second, the performance of the EDSN model with ECSC retained and EDSC removed decreases more significantly, with ACC and WF1 scores decreasing by 4.9% and 4.0%, respectively. These results show that ECSC and EDSC help to improve the accuracy of TER. The discriminative features extracted by EDSC have more emotional characteristics than the common features extracted by ECSC.

5.3.2 Comparative Analysis Between Modalities

To comprehensively assess the effects of different modalities on TER tasks, the ablation experiments are conducted by selecting and superimposing different modalities. The results are shown in Figure 8. First, only a single modality is selected as the input to EDSN. The first three columns (T, A, and V) of Figure 8 can be observed that the accuracy of the text-only modality (T) is as high as 0.699, which is significantly higher than the other two modalities. This phenomenon has been confirmed in other multimodal emotion recognition studies (Hazarika et al., 2020; Li et al., 2023; Mai et al., 2022). The possible reasons can be summarized as follows: First, the textual data is richer and less noisy, which contains rich emotional information; second, thanks to the powerful language models such as BERT,

Table 4 Evaluation of different components in EDSN

Model	Happiness		Satisfaction		Neutrality		Questioning		Total emotions	
	ACC	F1 score	ACC	F1 score	ACC	F1 score	ACC	F1 score	ACC	WF1 score
w/o ECSC	0.662	0.699	0.658	0.680	0.789	0.758	0.828	0.811	0.745	0.744
w/o EDSC	0.646	0.661	0.630	0.639	0.779	0.763	0.817	0.813	0.721	0.729

Note. w/o represents without.

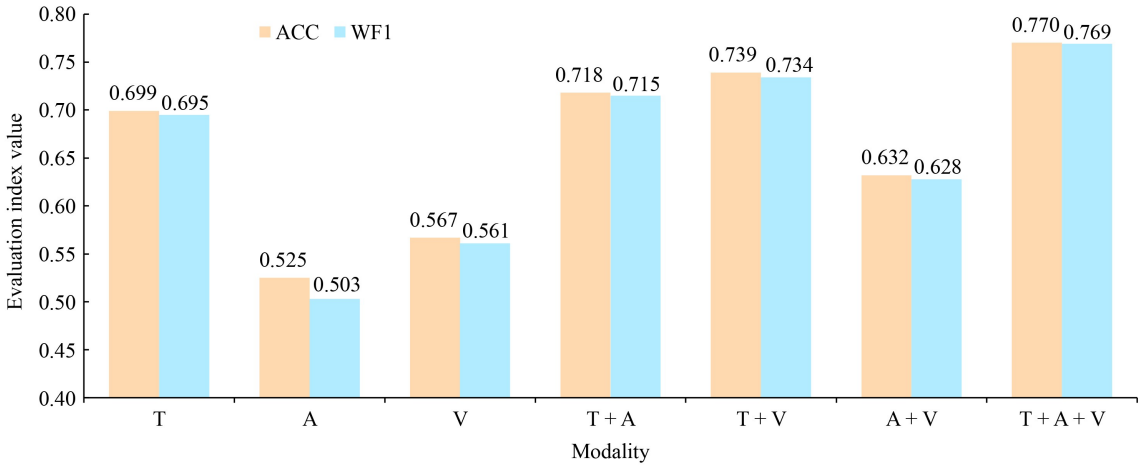

Figure 8 Comparison experiment of different input modalities on TER. T represents text-only modality, A represents audio-only modality, and V represents visual-only modality. ACC: accuracy; WF1: weighted F1 score.

the embedded emotional information is better characterized by understanding the contextual information. Additionally, two of the three modalities for arbitrary superposition are selected and find that the emotion recognition rate of two modalities superimposed is higher than that of a single modality, and that "T + V" outperforms "T + A" and "A + V", where T indicates text-only modality, V indicates video-only modality, and A means audio-only modality. This suggests that the superposition of textual and visual modalities maximizes the highlighting of teachers' emotion states. Three modalities are further developed, achieving the highest accuracy of 0.770, which suggests that each modality contributes to the TER task.

5.4 | Complexity and Convergence Evaluation

The number of parameters and their complexity are key factors affecting model performance, training efficiency, and generalization ability. A smaller number of parameters and lower computational complexity indicates a more lightweight model that is more suitable for deployment in resource-constrained environments. To provide a comprehensive evaluation of our EDSN, the number of parameters and computational complexity is analyzed in comparison with several baseline models as shown in Table 5. Experimental results show that while achieving more accurate emotion recognition accuracy, the proposed model has a smaller number of parameters and computational complexity compared to TFN, LMF, Graph-MFN, DMD, and MIB. This suggests that EDSN employs a simplified fusion method to efficiently extract common and discriminative emotional information from multimodal features, thereby significantly improving model performance while maintaining model simplicity.

Table 5 Comparative analysis of parameters and computational complexity of different models

Model	Number of training parameters ($\times 10^8$)	FLOPs ($\times 10^8$)
TFN	1.044	2.079
LMF	1.038	2.071
Graph-MFN	1.024	2.041
DMD	1.075	2.123
MIB	1.029	2.065
EDSN	**1.022**	**2.036**

Notes. The best results are highlighted in bold. FLOP: floating point operation.

To visualize the convergence rate of the models, the epoch–loss curve is ploted. As shown in Figure 9, the loss value of each model decreases as the training progresses and eventually converges. Moreover, the convergence speed of EDSN is significantly better than other models, and the loss value of EDSN is significantly lower than that of models such as TFN,

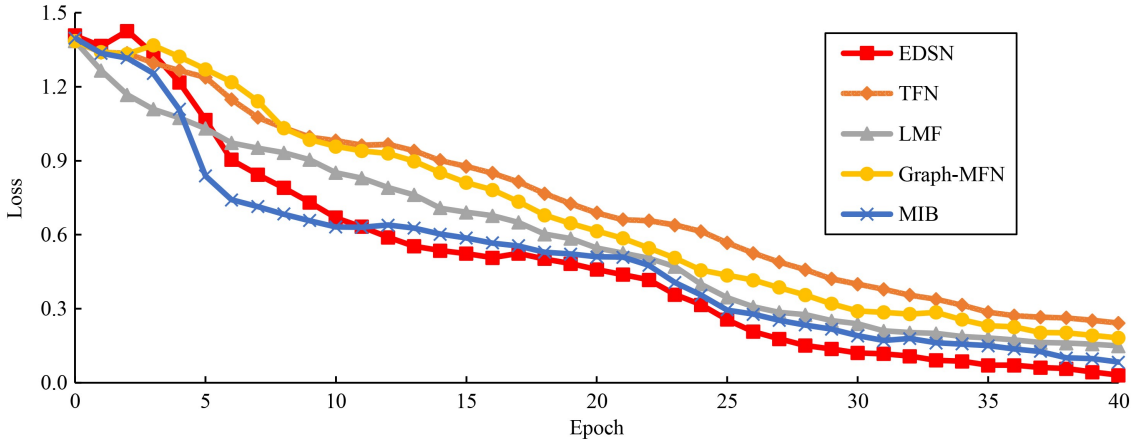

Figure 9 Convergence comparison of different models on TER dataset. EDSN: emotion dual-space network; TFN: tensor fusion network; LMF: low-rank multimodal fusion; Graph-MFN: graph memory fusion network; MIB: multimodal information bottleneck.

LMF, and Graph-MFN after 10 epoch value. The results demonstrate that EDSN not only converges faster, but also has a strong advantage in the final optimization results. This is attributed to its efficient extraction of multimodal features and keen ability to capture dynamic emotion changes. These characteristics enable EDSN to achieve superior performance more quickly and robustly under the same training conditions.

5.5 | Generalization Experiment

To further evaluate the generalization ability of the EDSN model, the experiment is conducted on the publicly available CMU-MOSI and IEMOCAP datasets.

5.5.1 Experiment on CMU-MOSI

The CMU-MOSI dataset is a commonly used benchmark dataset for evaluating the performance of multimodal emotion recognition models on the task of emotion intensity prediction. This dataset comprises numerous YouTube video blogs or vlogs, totaling 93 videos, each of which can be segmented into up to 62 utterances. It consists of 2,199 short monologue video clips, each manually annotated with a score ranging from −3 to +3, where −3 represents the strongest negative emotion and +3 denotes the strongest positive emotion. To facilitate comparison with the baseline model, several commonal used evaluation metrics are applied, such as mean absolute error (MAE), which measures the average absolute difference between predicted and true labels, seven-class classification accuracy (ACC 7 as shown in Table 6), which measures the proportion of predictions that match the true values in seven intervals ranging from −3 to +3, binary classification accuracy (ACC 2 as shown in Table 6), and F1 score computed for positive or negative classification outcomes.

Table 6 Comparison with other models on CMU-MOSI

Model	MAE (↓)	ACC 7 (↑)	ACC 2 (↑)	F1 score (↑)
TFN (Zadeh et al., 2017)	1.017	32.2	76.4	76.3
LMF (Liu et al., 2018)	1.026	30.6	73.8	73.7
MulT (Tsai et al., 2019)	1.009	33.6	79.3	78.3
PMR (Lv et al., 2021)	None	40.6	83.6	83.4
MISA (Hazarika et al., 2020)	0.783	42.3	83.4	83.6
CubeMLP (Sun et al., 2022)	0.770	45.5	85.6	85.5
MFSA (Yang et al., 2022b)	0.856	41.4	83.3	83.7
FDMER (Yang et al., 2022a)	0.724	44.1	84.6	84.7
MIB (Mai et al., 2022)	0.711	48.6	85.3	85.3
3MERNTLF (Zhao et al., 2024b)	0.907	None	80.8	80.9
DMD (Li et al., 2023)	None	45.6	**86.0**	**86.0**
TMRN (Lei et al., 2024)	**0.704**	**48.7**	85.7	85.5
Models in this research	0.725	48.2	**86.0**	**86.0**

Notes. The best results are highlighted in bold. The arrow "↓" means that the lower the data shows, the better the model performs. The arrow "↑" means that the higher the data shows, the better the model performs.

The experimental results on the CMU-MOSI dataset are shown in Table 6. Although the proposed model is designed specifically for recognition of TER, it still exhibits the best performance on the ACC 2 and F1 score on the CMU-MOSI dataset. This indicates that the proposed model achieves deep integration of text, audio, and visual modality by simultaneously considering both common and discriminative features of multimodal emotions, thereby enriching multimodal feature representation and demonstrating a degree of generalization capability. However, the proposed model does not achieve optimal performance on the MAE and ACC 7 metrics, indicating that there is room for improvement in regression and complex emotion recognition tasks. This may be due to the fact that the emotion categories in the CMU-MOSI dataset are more

nuanced, and that the differences in the different emotion labels are not only reflected in the emotional features, but also involve changes in the transitions between the features. This is a direction for continued research in the next stage.

5.5.2 Experiment on IEMOCAP

The IEMOCAP dataset is a widely used multimodal resource for emotion recognition research, encompassing textual, audio, and visual modalities. It comes from interactive conversations between multiple actors, with a total duration of about 12 hours, and contains 4,784 instances of spontaneous dialog, each labelled with emotional states such as happiness, sadness, anger and neutrality.

To validate the generalization ability of EDSN, the comparison experiments between EDSN and the latest baseline model at IEMOCAP are conducted. As shown in Table 7, the proposed EDSN achieves superior accuracy on the IEMOCAP dataset. Specifically, for the happy and angry labels, EDSN achieves an accuracy of 0.881 and 0.892 respectively, with F1 scores of 0.879 and 0.891, which is better than all other models. However, for sad and neutral emotional labels, EDSN is slightly inferior to the optimal model, but still maintains strong competitiveness. This may be because EDSN mainly focuses on the dynamic changes of emotion, while there are more emotion fragments with no obvious static and continuous changes in IEMOCAP dataset, which may make it difficult to fully exploit its advantages. Overall, EDSN can adapt to most multimodal datasets, which is also attributed to its comprehensive consideration of the common features and their respective discriminative features among multimodalities.

Table 7 Comparison with other models on IEMOCAP

Model	Happiness		Sadness		Anger		Neutrality	
	ACC	F1 score	ACC	F1 score	ACC	F1 score	ACC	F1 score
TFN	0.860	0.832	0.833	0.825	0.843	0.849	0.686	0.667
LMF	0.869	0.823	0.854	0.847	0.871	0.868	0.716	0.714
MulT	0.874	0.841	0.842	0.831	0.880	0.875	0.699	0.684
RMR	0.856	0.796	0.793	0.797	0.845	0.846	0.660	0.648
Graph-MFN	0.868	0.842	0.838	0.830	0.858	0.855	0.694	0.689
3MERNTLF	0.861	0.842	0.850	0.849	0.877	0.879	**0.731**	**0.729**
MIB	0.877	0.858	**0.876**	**0.869**	**0.892**	0.888	0.724	0.721
EDSN	**0.881**	**0.879**	0.856	0.854	0.886	**0.891**	0.723	0.722

Note. The best results are highlighted in bold.

6 Conclusions

This research first constructed a TER dataset suitable for real classroom situations. It not only supplements three kinds of modal information, but also inherits and extends the original emotion labels, and innovatively considers the teacher–student emotional interaction labels, including questioning and satisfaction. Given the difference in the emotional expressiveness of modalities on teachers, the research further proposed an EDSN to extract emotional common and discriminative features across multiple modalities respectively. To verify the generalization ability of the model, comparison experiments are also conducted on the public datasets CMU-MOSI and IEMOCAP. In summary, the TER dataset offers a robust foundation for future research on TER.

In the future, the data source will be gradually explanded by collecting more teaching videos from different languages and cultural backgrounds to enhance the diversity and generalization of the dataset. Meanwhile, more abundant emotional labels related to classroom scenes will be further marked to improve the data expression and applicability. In terms of model design, the practical deployment of EDSN in real classrooms can be further explored, and the impact of other interfering factors on data quality can be analyzed in depth to provide support for improving the accuracy and utility of TER.

Nomenclature

ACC	Accuracy
BERT	Bidirectional encoder representations from transformers
Conv1D	1-dimensional convolutional layer
CMD	Central moment discrepancy
DMD	Decoupled multimodal distillation
ECSC	Emotion commonality space construction
EDSC	Emotion discrimination space construction
EDSN	Emotion dual-space network
FLOP	Floating point operation
Graph-MFN	Graph memory fusion network
GRL	Gradient reversal layer
LMF	Low-rank multimodal fusion
MAE	Mean absolute error
MAG	Multimodal adaptation gate
MFN	Multimodal fusion network
MIB	Multimodal information bottleneck
MLP	Multilayer perceptron
MOOC	Massive open online course
MTED	Multimodal teacher emotion dataset
TER dataset	Multimodal teacher emotion recognition dataset

TFN	Tensor fusion network
WF1	Weighted F1

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant Nos. 62377007 and 62407009), the Chongqing University Graduate Education Teaching Reform Research Key Project, China (Grant No. 232073), the Scientific and Technological Research Program of Chongqing Municipal Education Commission, China (Grant Nos. KJZD-M202400606 and KJZD-M202300603), and the Chongqing Natural Science Foundation Joint Key Project for Innovation and Development, China (Grant No. 2024NSCQ-LZX0057).

Authors Contributions Ting Cai made substantial contributions to the conception and design of the work; Shengsong Wang analyzed the data; Jing Wang interpreted the data; Yu Xiong drafted the work; Long Liu revised it critically for important intellectual content. All authors approved the version to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflict of Interest The authors declare that they have no conflict of interest.

Ethics Statements The authors confirm that their Institutional Ethics Committee determined that no ethical review was required for this study. Written informed consent for participation was not required because all participant data were anonymized prior to statistical analysis.

Data Availability Statements The TER dataset is designed for scholarly research on teachers' emotion recognition. It is intended for noncommercial, scholarly use only, and proper citation of this article and acknowledgement of the dataset source is required in any derivative works or publications. Access can be obtained by contacting the corresponding author. The dataset has been anonymized to protect the privacy of the participants, and researchers must adhere to ethical standards to ensure no misuse of the data. For questions, please contact the corresponding author.

References

Cai, T., Xiong, Y., He, C. Y., Wu, C., & Zhou, S. (2024). TBU: A large-scale multi-mask video dataset for teacher behavior understanding. In: *Proceedings of 2024 IEEE International Conference on Multimedia and Expo*. Niagara Falls: IEEE, 1–6.

Department of Education of the Republic of the Philippine. (2019, Feburary). *Code of ethics for professional teachers*. Available from UNESCO website.

Dreer, B. (2024). Teachers' well-being and job satisfaction: The important role of positive emotions in the workplace. *Educational Studies*, *50*(1), 61–77.

Ekman, P. (1984). Expression and the nature of emotion. In: Scherer, K. R., & Ekman, P., eds. *Approaches to emotion*. New York: Psychology Press, 319–343.

Frenzel, A. C., Daniels, L., & Burić, I. (2021). Teacher emotions in the classroom and their implications for students. *Educational Psychologist*, *56*(4), 250–264.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In: *Proceedings of the 32nd International Conference on Machine Learning*. Lille: JMLR.org, 1180–1189.

Geetha, A. V., Mala, T., Priyanka, D., & Uma, E. (2024). Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Information Fusion*, *105*, 102218.

Han, W., Chen, H., Gelbukh, A., Zadeh, A., Morency, L. P., & Poria, S. (2021). Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In: *Proceedings of the 2021 International Conference on Multimodal Interaction*. Montréal: ACM, 6–15.

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In: *Proceedings of the 28th ACM International Conference on Multimedia*. Seattle: ACM, 1122–1131.

Jiang, J. W., Vauras, M., Volet, S., & Wang, Y. L. (2016). Teachers' emotions and emotion regulation strategies: Self- and students' perceptions. *Teaching and Teacher Education*, *54*, 22–31.

Keller, M. M., & Becker, E. S. (2021). Teachers' emotions and emotional authenticity: Do they matter to students' emotional responses in the classroom? *Teachers and Teaching*, *27*(5), 404–422.

Khan, U. A., Xu, Q. R., Liu, Y., Lagstedt, A., Alamäki, A., & Kauttonen, J. (2024). Exploring contactless techniques in multimodal emotion recognition: Insights into diverse applications, challenges, solutions, and prospects. *Multimedia Systems*, *30*(3), 115.

Lei, Y. X., Yang, D. K., Li, M. C., Wang, S. L., Chen, J. W., & Zhang, L. H. (2024). Text-oriented modality reinforcement network for multimodal sentiment analysis from unaligned multimodal sequences. In: *Proceedings of the Artificial Intelligence: Third CAAI International Conference*. Fuzhou: Springer, 189–200.

Li, Y., Wang, Y. Z., & Cui, Z. (2023). Decoupled multimodal distilling for emotion recognition. In: *Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver: IEEE, 6631–6640.

Liang, J., Zhao, X. Y., & Zhang, Z. H. (2020). Speech emotion recognition of teachers in classroom teaching. In: *Proceedings of 2020 Chinese Control and Decision Conference*. Hefei: IEEE, 5045–5050.

Liu, S. H., Wang, Y. X., Wang, K. H., Li, B. S., Yang, F. Q., & Yang, S. H. (2024). Semantic-wise guidance for efficient multimodal emotion recognition with missing modalities. *Multimedia Systems*, *30*(3), 144.

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A. A. B., & Morency, L.-P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne: ACL, 2247–2256.

Lu, Y. Y., Chen, Z. Z., Zheng, Q. Y., Zhu, Y. H., & Wang, M. K. (2023). RETRACTED ARTICLE: Exploring multimodal data analysis for emotion recognition in teachers' teaching behavior based on LSTM and MSCNN. *Soft Computing*, *28*(2), 699.

Lv, F. M., Chen, X., Huang, Y. Y., Duan, L. X., & Lin, G. S. (2021). Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In: *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2554–2562.

Mai, S. J., Zeng, Y., & Hu, H. F. (2022). Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, *25*, 4121–4134.

McFee, B., Raffel, C., Liang, D. W., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. In: *Proceedings of the 14th Python in Science Conference*. 18–24.

Ong, S. G. T., & Quek, G. C. L. (2023). Enhancing teacher–student interactions and student online engagement in an online learning environment. *Learning Environments Research*, *26*(3), 681–707.

Qin, Q., Hu, W. P., & Liu, B. (2020). Feature projection for improved text classification. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 8161–8171.

Rahman, W., Hasan, M. K., Lee, S., Zadeh, B. A. A., Mao, C. F., Morency, L.-P., & Hoque, E. (2020). Integrating multimodal information in large pretrained transformers. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 2359–2369.

Sun, H., Wang, H. Y., Liu, J. Q., Chen, Y. W., & Lin, L. F. (2022). CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. In: *Proceedings of the 30th ACM International Conference on Multimedia*. Lisboa: ACM, 3722–3729.

Tsai, Y. H. H., Bai, S. J., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: ACL, 6558–6569.

Uzuntiryaki-Kondakci, E., Kirbulut, Z. D., Sarici, E., & Oktay, O. (2022). Emotion regulation as a mediator of the influence of science teacher emotions on teacher efficacy beliefs. *Educational Studies*, *48*(5), 583–601.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 6000–6010.

Wang, S. Y., Cheng, L. M., Liu, D. Y., Qin, J. Q., & Hu, G. H. (2022). Classroom video image emotion analysis method for online teaching quality evaluation. *Traitement du Signal*, *39*(5), 1767–1774.

Yang, D. K., Huang, S., Kuang, H. P., Du, Y. T., & Zhang, L. H. (2022a). Disentangled representation learning for multimodal emotion recognition. In: *Proceedings of the 30th ACM International Conference on Multimedia*. Lisboa: ACM, 1642–1651.

Yang, D. K., Kuang, H. P., Huang, S., & Zhang, L. H. (2022b). Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences. In: *Proceedings of the 30th ACM International Conference on Multimedia*. Lisboa: ACM, 1708–1717.

Yang, Y., Gong, S. Y., Cao, Y., Qiu, Y., Xu, X. Z., & Wang, Y. Q. (2024). Linking teacher support to achievement emotion profile: The mediating role of basic psychological need satisfaction. *Frontiers in Psychology*, *15*, 1352337.

Yu, W. M., Xu, H., Meng, F. Y., Zhu, Y. L., Ma, Y. X., Wu, J. L., Zou, J. Y. & Yang, K. C. (2020). CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 3718–3727.

Zadeh, A., Chen, M. H., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen: ACL, 1103–1114.

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L.-P. (2018). 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: ACL, 2236–2246.

Zadeh, A., Zellers, R., Pincus, E., & Morency, L.-P. (2016). MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv Preprint*, arXiv:1606.06259.

Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., & Saminger-Platz, S. (2017). Central moment discrepancy (CMD) for domain-invariant representation learning. In: *Proceedings of the 5th International Conference on Learning Representations*. Toulon: OpenReview.net.

Zhang, J. H., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, *59*, 103–126.

Zhao, G., Zhang, Y. N., & Chu, J. (2024a). A multimodal teacher speech emotion recognition method in the smart classroom. *Internet of Things*, *25*, 101069.

Zhao, L. L., Yang, Y. L., & Ning, T. (2024b). A Three-stage multimodal emotion recognition network based on text low-rank fusion. *Multimedia Systems*, *30*(3), 142.