



Semantic-wise guidance for efficient multimodal emotion recognition with missing modalities

Shuhua Liu¹ · Yixuan Wang¹ · Kehan Wang¹ · Binshuai Li¹ · Fengqin Yang¹ · Shihao Yang¹

Received: 26 August 2023 / Accepted: 3 March 2024 / Published online: 9 May 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Emotions play an important role in human–computer interaction. Multimodal emotion recognition combines feature information from different modalities to recognize emotional states. However, in real application scenarios, data from all modalities may not always be available. Thus, in multimodal emotion recognition a big challenge is how to utilize the semantic information from available modalities to predict missing modality data. To address this issue, this study proposes a Semantic-Wise Guidance for Missing Modality Imagination Network (SWG-MMIN) consisting of three main modules, that is, the Comprehensive Modality Feature Enrichment (CMFE) module, the Semantic-Wise Fusion (SWF) module, and the Semantic-Wise Feature Guided Imagination (SWGFI) module. The CMFE module addresses the issue of semantic loss in the process of integrating multimodal features by enhancing the semantic information. The SWF module performs an adaptive fusion of invariant and specific features of multimodal data. The SWGI module facilitates the missing modality data generation and enhances the robustness of joint multimodal representation. Extensive experiments are conducted on two benchmark datasets, IEMOCAP and MSP-IMPROV. The experimental results demonstrate that the SWG-MMIN model surpasses all baseline models under full modalities and uncertain missing modalities, significantly improving emotion recognition performance.

Keywords Emotion recognition · Modality feature enrichment · Semantic-wise fusion · Missing modality imagination

Communicated by X. Li.

✉ Shihao Yang
yangsh861@nenu.edu.cn

Shuhua Liu
liush129@nenu.edu.cn

Yixuan Wang
wangyx504@nenu.edu.cn

Kehan Wang
wangkh186@nenu.edu.cn

Binshuai Li
libs429@nenu.edu.cn

Fengqin Yang
yangfq147@nenu.edu.cn

¹ School of Information Science and Technology, Northeast Normal University, Changchun, China

1 Introduction

Multimodal emotion recognition effectively performs emotion recognition tasks by integrating feature information from different modalities [1, 2, 28]. However, in real-world scenarios, missing modalities are common due to environmental factors [17], such as audio loss or camera damage, leading to insufficient data [6], as shown in Fig. 1. Traditional approaches typically address the issue of missing modalities through data generation methods [30] or joint multimodal representation learning methods [26, 33]. Zhao et al. [37] proposed Missing Modality Imagination Network (MMIN) for emotion recognition with uncertain missing modalities, which integrated the above two methods to build a network for addressing emotion recognition with uncertain missing modalities. The effectiveness of MMIN surpasses that of using a single method. However, this approach can be improved in the following areas:

Semantic loss in the process of unifying the dimension of multimodal features. In MMIN [37], before multimodal feature fusion, the features of different modalities are reduced



Modality	Content	Predict
Acoustic		
Visual		Neutral
Textual	In fact, I don't like rainy days, I like sunny days.	

Fig. 1 Case of two missing modalities, where missing modalities are marked with dotted red lines

dimension by their respective modality encoders to obtain unified dimension. However, the dimension reduction process inevitably results in the loss of certain semantic information. This implies that important semantic information may be lost during this process, which leads to a decrease in the recognition accuracy of the model. Specifically, when dealing with the task of missing modalities, the insufficient feature information of the available modality hinders the module's imagination ability to accurately generate the missing modalities. And this leads to a substantial discrepancy between the generated missing modalities and the real modalities. Thus, this discrepancy impacts the overall recognition performance of the model.

Weight assignment between modality-invariant and modality-specific features in multimodal feature fusion. Hazarika et al. [12] proposed the Modality-Invariant and Modality-Specific Representations for Multimodal Sentiment Analysis (MISA) model, which projects each modality feature into a shared semantic space to identify potential commonalities among modalities. This approach alleviates the inherent discrepancies across heterogeneous modalities. However, MISA fails to address the appropriate weights assignment to the joint representation of invariant and specific features across modalities. For example, some emotions are more easily recognized by language, while others are more easily recognized by vision. Therefore, this study proposes that the joint representation of modality-invariant and modality-specific features should be assigned distinct weights in multimodal fusion. Otherwise, it will result in suboptimal joint multimodal representation and cause a decrease in recognition accuracy.

Aiming at the issues of semantic loss in the process of dimension reduction and suboptimal joint multimodal representation, this study proposes the Semantic-Wise Guidance for Missing Modality Imagination Network (SWG-MMIN). Firstly, the Comprehensive Modality Feature

Enrichment module is introduced to obtain richer semantic information on the modality invariant features and specific features. Next, the Semantic-Wise Fusion module is employed to adaptively fuse the obtained invariant features and specific features, so as to realize the adaptive joint representation of multimodal data. Finally, the fused features are fed into the Semantic-Wise Feature Guided Imagination module to help generate missing modality data and improve the robustness of the joint multimodal representation. The experimental results on two benchmark datasets of IEMOCAP and MSP-IMPROV show that the SWG-MMIN model outperforms other advanced models under the conditions of full modalities and uncertain missing modalities.

The contributions of this work can be summarized as follows:

1. This paper proposes an emotion recognition model Semantic-Wise Guidance for Missing Modality Imagination Network (SWG-MMIN), which contains the Comprehensive Modality Feature Enrichment module, the Semantic-Wise Fusion module, and the Semantic-Wise Feature Guided Imagination module. The effectiveness of the proposed model is verified on the widely used IEMOCAP and MSP-IMPROV datasets. The experimental results demonstrate that the SWG-MMIN model achieves significant performance improvement compared to other models under the conditions of full modalities and uncertain missing modalities.
2. This paper introduces the Comprehensive Modality Feature Enrichment module that enriches the semantic information of multimodal features, addressing the issue of semantic loss in the process of unifying multimodal features. By reducing the semantic loss, the model obtains modality-specific and modality-invariant features with richer semantic information, thereby improving the overall recognition performance of the model.
3. The Semantic-Wise Fusion module is designed to perform an adaptive fusion of invariant and specific features in multimodal features. This module assigns appropriate weights to modality-invariant and modality-specific features, enabling the model to learn a robust joint multimodal representation during cross-modality imagination.

2 Related work

Multimodal emotion recognition. The current methods for multimodal emotion recognition (MER) [39] can be divided into two categories, that is, fusion strategy-based

methods [12, 21, 35] and cross-modality attention-based methods [18, 23, 31]. (1) Fusion strategy-based methods aim to design complex fusion strategies to generate robust multimodal representations and improve emotion recognition performance. Zadeh et al. [35] proposed the Tensor Fusion Network (TFN) which is an end-to-end multimodal sentiment analysis model. TFN implements dynamic modeling of intra-modality and inter-modality through tensor fusion and modality embedding sub-networks. However, its high dimension and exponential computational complexity limit its further development and application. Liu et al. (Z. [21]) proposed a low-rank multimodal fusion method that decomposes tensors and weights. It employs modality-specific low-rank factors for multimodal fusion so as to avoid the calculation of high-dimensional tensor and reduce memory consumption. However, this method does not consider modality-invariant features. Hazarika et al. [12] emphasized the significance of representation learning in multimodal fusion. They proposed a method that projects each modality into modality-invariant and modality-specific spaces to learn a more comprehensive multimodal representation. However, this study did not consider the adaptive weight assignment for invariant and specific features. (2) Methods based on Cross-modality attention aim to learn inter-modality correlations to obtain robust multimodal representations. Tsai et al. [31] proposed Multimodal Transformer (MulT) consisting of paired cross-modality attention mechanisms. This model provides a potential cross-modality adaptation by directly focusing on low-level features of other modalities. However, the problem of multimodal alignment is still a challenge in practical applications. Liang et al. [18] proposed a new semi-supervised multimodal emotion recognition model (SSMM) based on cross-modality distribution matching. This method utilizes an amount of unlabeled data to train model and improve emotion recognition performance. However, it is a multimodal fusion model trained based on full-modality samples and its performance decreases greatly in the absence of partial modalities. Therefore, this study proposes the SWG-MMIN model, which comprehensively and effectively utilizes the specific and invariant feature information from different modalities. The Semantic-Wise Fusion module is employed to adaptively fuse these features, enabling the model to learn a robust multimodal joint representation. Thus, the SWG-MMIN model performs accurate and efficient emotion recognition in scenarios under full and missing modalities due to overcome the limitations of other methods.

Modality feature enhancement. Feature enhancement techniques are widely adopted in computer vision (J. [14, 15, 19]). Huang et al. [15] proposed a densely connected

convolutional network (DenseNet). By introducing dense connections, the feature maps of each layer can be connected with subsequent layers so as to enhance the ability to transfer and reuse features. Lin et al. [19] proposed the Feature Pyramid Network (FPN). By introducing multi-scale feature pyramids, the network can obtain feature information of different scales, thereby improving the performance of the object detection task. In addition, some research methods combine visual and textual information so that the textual features can get knowledge from the visual modality [24, 36]. Mun et al. [24] proposed a text-guided attention model. The model enhances the attention to image features and the accuracy of description generation by introducing textual information into the attention mechanism in image description generation. Zhang et al. [36] adopted the sentence embedding framework SimCSE and extended it to a multimodal contrastive objective. At the same time, they utilized both visual and text information, where visual information serves as auxiliary semantic information to further performance sentence representation learning. However, in multimodal emotion recognition, the semantic information contained in different modality features is crucial to the performance of emotion recognition. Therefore, it is necessary to further explore how to enhance different modality features and reduce semantic loss when unifying multi-modality. To this end, this study introduces the Comprehensive Modality Feature Enrichment module, which aims to obtain feature information of different scales to compensate for semantic loss in the dimension reduction of multimodal features. Owing to this module, richer emotional features can be obtained, and then improving the performance of emotion recognition.

Learning joint multimodal representations. In recent years, research on the missing modality issue has mainly focused on methods for learning joint multimodal representations [11, 27, 32, 33], Zhao, Li, Jin, et al. [37] to encode all modality information. Han et al. [11] proposed a model that implicitly fuses audio and video information during the training of speech or facial emotion recognition. However, this model only enhances text modality emotion recognition and does not fully utilize the semantic relationships between modalities. Wang et al. [33] proposed a Transformer-based translation model, which simulates the relationship between the source and target languages by using emotional mechanisms. The model employs a parallel translation approach to fuse text features with acoustic features, text features with visual features. The model also adopts forward and backward translation strategies to better fuse multimodal features. However, this model cannot handle emotion recognition scenarios under uncertain missing modalities, and different models need to be established for

different missing modality conditions. Therefore, this study introduces the Semantic-Wise Feature Guided Imagination module. This module uses the adaptively fused feature of available modalities and cascades them into each auto-encoder to assist generating the imagination data of missing modalities. By this process, the model can learn a robust joint multimodal representation. Therefore, the module is able to effectively handle the emotion recognition scenarios with various missing modalities.

3 Method

3.1 Overview

This study proposes a Semantic-Wise Guidance for Missing Modalities Imagination Network (SWG-MMIN). The model can effectively and comprehensively extract the semantic information from modality specific and invariant features to imagine missing modalities through efficient semantic guidance. By adaptively fusing modality features, the proposed model learns a robust joint multimodal representation and achieves accurate emotion recognition in scenarios under both full modalities and missing modalities. The overall framework of the model is shown in Fig. 2.

Given a set of video segments, $x = (x^a, x^v, x^t)$. represents the raw multimodal features, where x^a, x^v and x^t represent the raw features of acoustic, visual, and textual modalities,

respectively. First, a Comprehensive Modality Feature Enrichment module and a Semantic-Wise Fusion module are introduced under full modalities to pretrain the specificity encoder and invariance encoder. And then SWG-MMIN is further trained under missing modalities to form a complete model framework. As shown in Fig. 2, the visual modality is missing and represented by x^v_{miss} , and then the input of the model is represented as (x^a, x^v_{miss}, x^t) . First, the above triplet is input into the Comprehensive Modality Feature Enrichment module to enhance the features of each modality. It can reduce the semantic loss of the modality due to dimension reduction. Second, the Semantic-Wise Fusion module is employed to adaptively fuse the invariant features (H^a, H^v_{miss}, H^t) and the specific features (h^a, h^v_{miss}, h^t) of the heterogeneous modality. The adaptively fused weighted features h_{fusion} contain both invariant and specific features with different weights. Finally, h_{fusion} is input into the Semantic-Wise Feature Guided Imagination module to help predict the missing modality embeddings. The latent vectors of each auto-encoder in the imagination module are collected and concatenated to form a joint multimodal representation S , which is then input into the classifier for emotion recognition. The pretrained Specificity and Invariance Encoder in the full-modality scenario are similar to the Specificity and Invariance Encoder in the SWG-MMIN training process, with the only difference being that the parameters of the pretrained Specificity and

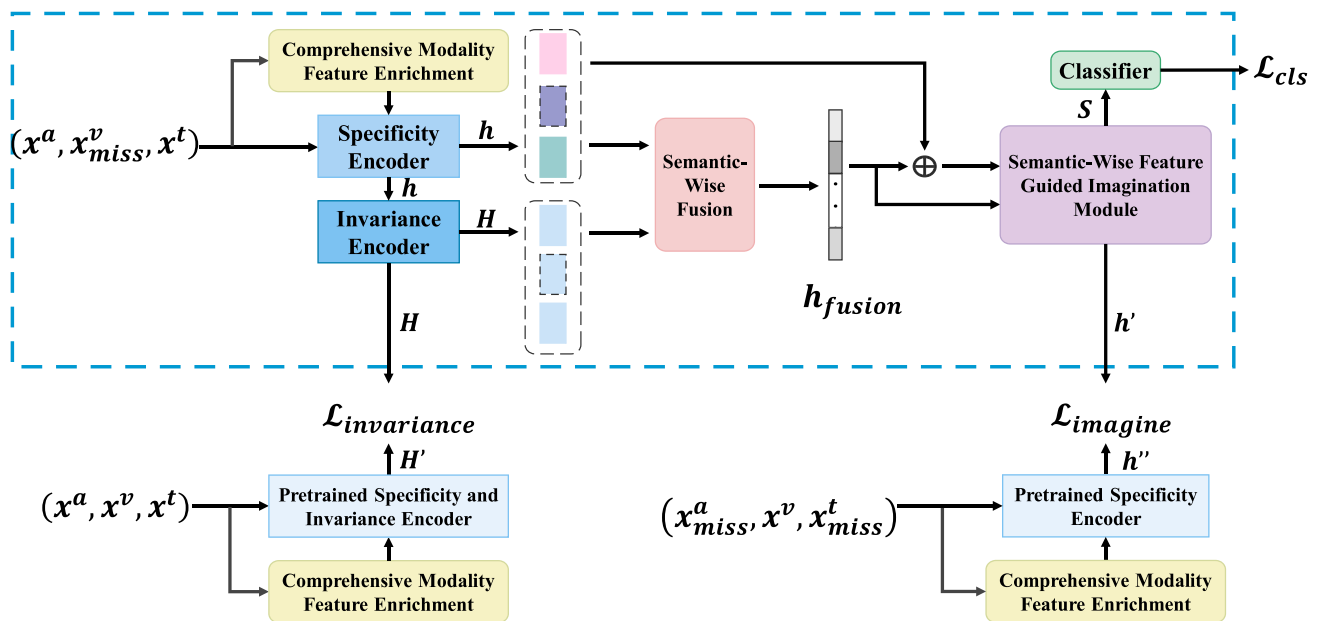


Fig. 2 The framework of the proposed SWG-MMIN. SWG-MMIN is trained with all six possible missing modality conditions. Taking missing visual modality as an example, the modality-specific feature h is represented as $h = (h^a, h^v_{miss}, h^t)$ and the modality-invariant H is

represented as $H = (H^a, H^v_{miss}, H^t)$. The Pretrained Specificity and Invariance Encoder, as well as the Pretrained Specificity Encoder, are pretrained under full-modality data, and the parameters of two blocks remain fixed during the SWG-MMIN training process

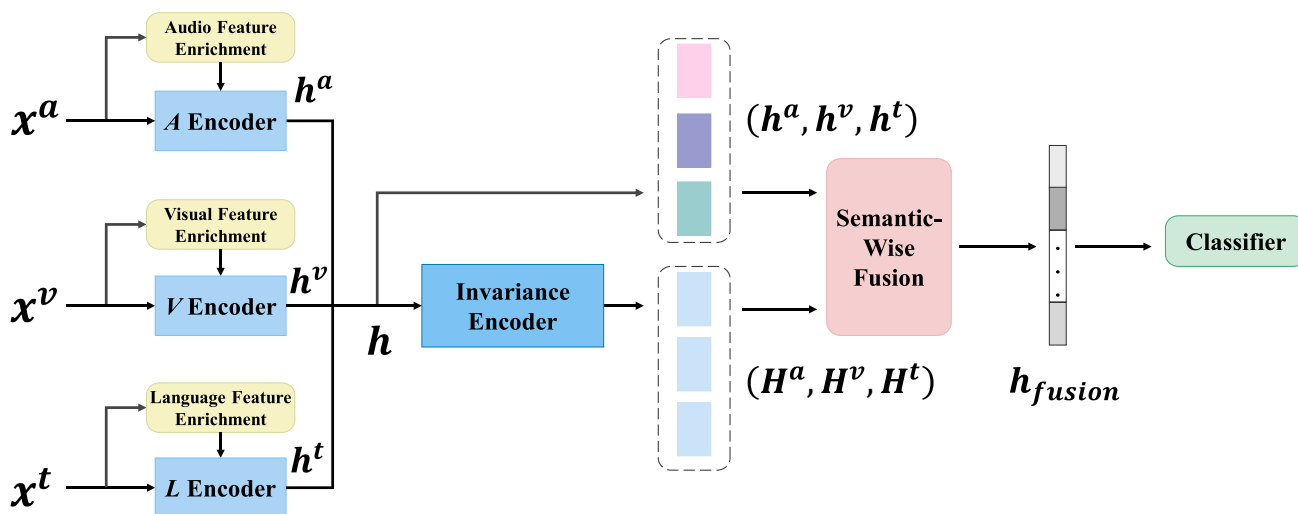


Fig. 3 Pretrained Encoder network under full modalities based on Semantic-Wise Guidance

Invariance Encoder remain fixed during the SWG-MMIN training process.

3.2 Pretrained encoder network under full modalities based on semantic-wise guidance

Figure 3 shows the pretrained network for each specificity encoder and invariance encoder under full modalities.

As shown in Fig. 3, the pipeline based on Semantic-Wise Guidance network learning mainly contains the following modules: the Comprehensive Modality Feature Enrichment (CMFE) module, the modality specificity encoders, invariance encoder, the Semantic-Wise Fusion (SWF) module, and the classifier. The CMFE module aims to enhance the semantic information of the specificity encoder in extracting modality-specific high-level features $h = (h_0^a, h_0^v, h_0^t)$ from the raw features $x = (x^a, x^v, x^t)$. Specifically, the Acoustic Encoder (*A Encoder*) employs an LSTM network and a max-pooling layer to extract utterance-level acoustic features $h_0^a = AEncoder(x^a)$. The enhanced acoustic features are represented as h^a . The Visual Encoder (*V Encoder*) employs a structure similar to the

Acoustic Encoder to extract features $h_0^v = VEncoder(x^v)$. After the enhancement, the visual features are represented as h^v . The text encoder (*L Encoder*) employs Text CNN to extract utterance-level text features $h_0^t = LEncoder(x^t)$. The enhanced text features are represented as h^t . The invariance encoder takes (h^a, h^v, h^t) as input and maps the cross-modality features to a shared subspace through a learning strategy based on the central moment discrepancy (CMD) distance. This strategy reduces the difference between the shared representations of each modality by minimizing the CMD loss to obtain modality-invariant features (H^a, H^v, H^t) . Finally, the modality-specific features (h^a, h^v, h^t) and the modality-invariant features (H^a, H^v, H^t) are input into the SWF module for adaptive fusion to obtain the Semantic-Wise feature, which is then input to the classifier for multimodal emotion recognition.

Next, the Comprehensive Modality Feature Enrichment (CMFE) module and the Semantic-Wise Fusion module (SWF Module) are introduced in details.

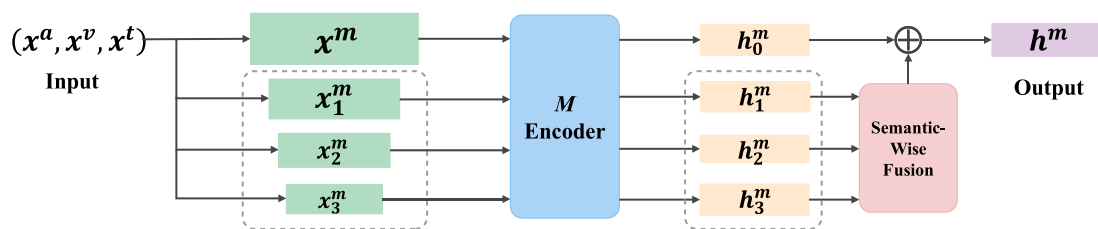


Fig. 4 Schematic diagram of the Comprehensive Modality Feature Enrichment Module, where M Encoder denotes the modality-specific encoder for different modalities, $M \in A, V, L$

3.2.1 Comprehensive modality feature enrichment module

Before multimodal fusion, it is necessary to encode heterogeneous modalities and transform the raw features into low-dimensional utterance-level representations to ensure unified dimensions for subsequent joint representations. However, semantic loss will inevitably occur in the process of dimension reduction for heterogeneous modalities. That is, these incomplete modality features further deepen the semantic gap between heterogeneous modalities. This will bring a negative impact on multimodal joint representation and reduce the recognition accuracy. To address the semantic loss issue caused by dimension reduction, this study introduces a Comprehensive Modality Feature Enrichment (CMFE) module, as shown in Fig. 4.

Specifically, this study adopts the Avgpool operation to downsample the raw features $x^m(m \in a, v, t)$ with downsampling coefficients of 2, 4, and 8, respectively. Thus, downsampled features at three different scales x_1^m, x_2^m, x_3^m are obtained. Next, the downsampled features are input into the same modality-specificity encoder as the raw features, and four features $h_0^m, h_1^m, h_2^m, h_3^m$ with the same dimension are obtained. Among them, h_0^m represents the modal-specificity features obtained from the raw features by the encoder and h_1^m, h_2^m, h_3^m represents the modality semantic enrichment features obtained from the downsampling operation by the encoder. Next, h_1^m, h_2^m, h_3^m are fused through a Semantic-wise Fusion module, and the fused semantic enrichment feature is combined with h_0^m to obtain feature h^m . Compared to the feature h_0^m obtained from the raw features by the encoder, feature h^m contains richer multi-scale semantic information, achieving the objective of feature semantic enhancement. The complete process can be expressed as follows:

$$h^m = \text{Encoder}(x^m) + \text{SWF}(\text{Encoder}(\text{Avgpool}_{2^i}(x^m)), \text{Encoder}(\text{Avgpool}_{2^{i+1}}(x^m)), \text{Encoder}(\text{Avgpool}_{2^{i+2}}(x))) \quad (1)$$

where Encoder represents the corresponding modality encoder and SWF represents Semantic-wise Fusion module, and Avgpool $_{\gamma}$ represent Avgpool with downsampling factor $\gamma(\gamma = 2, 4, 8)$.

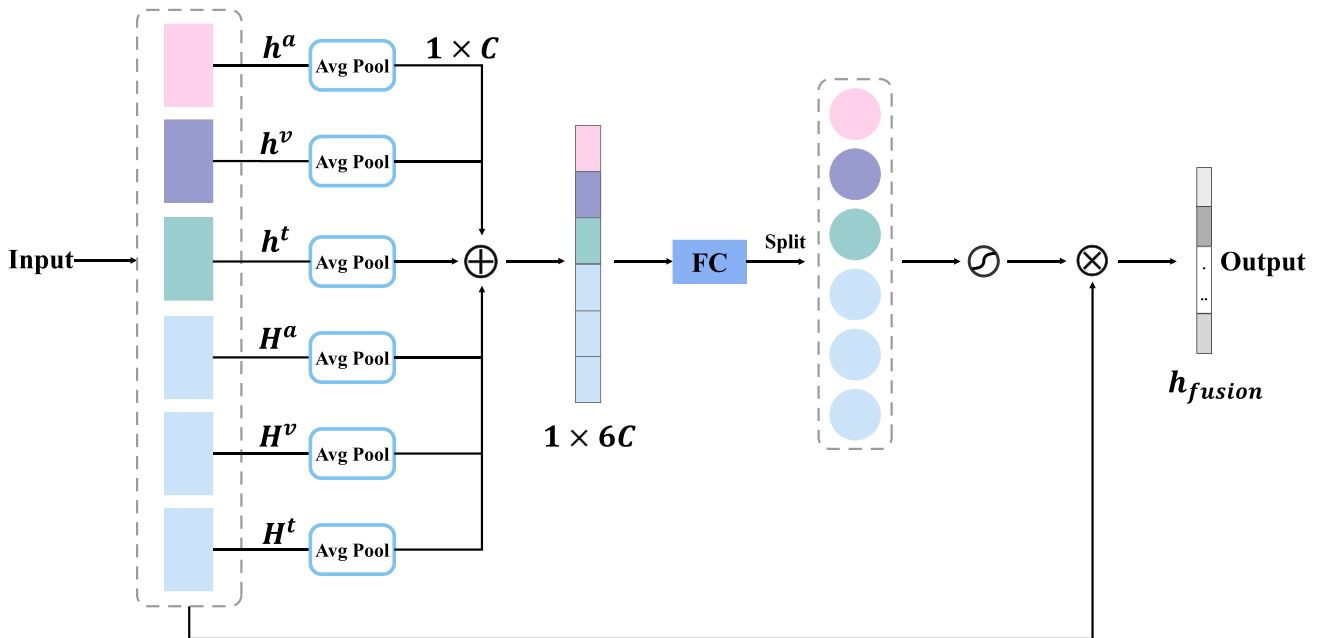


Fig. 5 Schematic diagram of Semantic-wise Fusion module, where C represents the output size of the unified multimodal feature dimension after the encoder

3.2.2 Semantic-wise fusion module

In the joint representation of multimodal emotion recognition, different modalities have both common and unique semantic characteristics in the semantic space. The invariant features of heterogeneous modalities reflect the shared motives and goals expressed by each modality and these invariant features contribute to the understanding of global emotional states. Meanwhile, the specific features of heterogeneous modalities emphasize the distinctive emotions expressed by each modality. If their joint representation without considering weights, it may result in the model failing to handle the interrelationship between invariant and specific features across heterogeneous modalities. This will lead to a negative impact on emotion recognition tasks. To address this issue, this study proposes a Semantic-Wise Fusion (SWF) module to enable adaptive weight assignment between modality invariant and specific features, thereby obtaining an efficient joint multimodal representation. The structure of the module is shown in Fig. 5.

The specific and invariant features $h^a, h^v, h^t, H^a, H^v, H^t$ of heterogeneous modalities are used as input. And global Avg-pool operation is performed on each of the above features, respectively, to obtain the global context information with scale $1 \times C$. Then the global context information is fused to obtain the fusion feature with scale $1 \times 6C$. The fusion feature is input into an FC layer, and then the output of the FC layer is split into 6 contexts with scale $1 \times C$. In the following, the 6 contexts are activated by the sigmoid function to obtain the fusion weights $V_1, V_2, V_3, V_4, V_5, V_6$ of the 6 different features. Finally, the fusion weights multiply with the corresponding input $h^a, h^v, h^t, H^a, H^v, H^t$ and obtain h_{fusion} with adaptive fusion weights. This design fully utilizes the invariant features and specific features to achieve efficient multimodal feature fusion to implement more accurate emotion recognition. The calculation process of joint multimodal representation h_{fusion} with adaptive fusion weights is represented as formula (2).

$$h_{fusion} = h^a \times V_1 + h^v \times V_2 + h^t \times V_3 + H^a \times V_4 + H^v \times V_5 + H^t \times V_6 \tag{2}$$

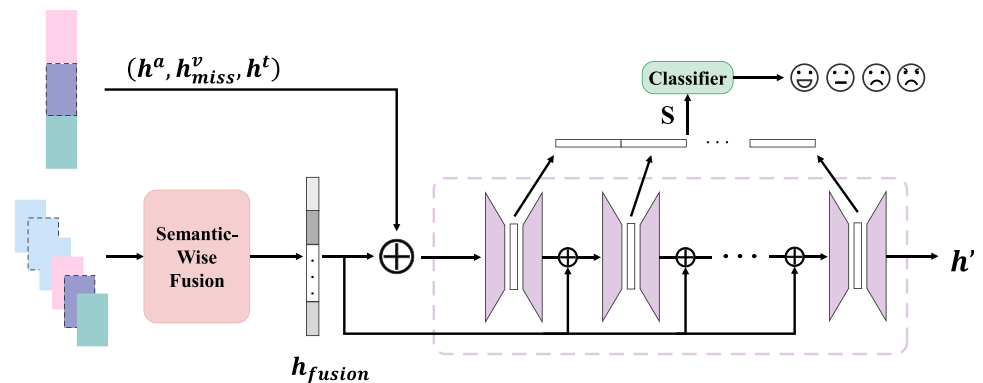
3.3 Semantic-wise feature guided imagination module under missing modalities

To ensure the stability and accuracy of emotion recognition under missing modalities, this study introduces the Semantic-Wise Feature Guided Imagination (SWGFI) module in the construction of the SWG-MMIN model. The SWGI module enables the SWG-MMIN model to imagine missing modality data under different missing-modality conditions. The structure of the SWGI module is shown in Fig. 6.

The SWGI module employs a cascaded auto-encoder structure containing N auto-encoders to predict the multimodal embeddings of missing modalities based on the available modalities. Different from MMIN (Zhao, Li and Jin [37]), which directly inputs modality-specific features after concatenation, the SWGI module also reads the available modality adaptive fusion feature h_{fusion} and cascaded input it into each auto-encoder. The fused feature involves both the unique semantics of specific features and the shared semantics of invariant features, assisting the generation of missing modality data and alleviating the gap among heterogeneous modalities.

Taking the missing visual modality as an example, represented by h^v_{miss} , the multimodal embeddings of across-modality pairs under the missing condition is represented as $h_c = \text{concat}(h^a, h^v_{miss}, h^t)$. The fused feature of specific and invariant features of the available modality is expressed as h_{fusion} . Features h_c adds h_{fusion} and the result denoted as h_F , which is to be fed into the SWGI module for multimodal imagination. Additionally, h_{fusion} also being cascaded input into each auto-encoder to assist in generating missing modality data. Each auto-encoder learns from available modality data to acquire latent representations. It utilizes correlations between different modalities to reconstruct and estimate the missing modality data. Simultaneously setting imagination loss to make the output data of forward learning closer to

Fig. 6 Detailed structure of Semantic-Wise Feature Guided Imagination module



real data. Therefore, the SWGI module is a cascaded auto-encoder model consisting of N auto-encoders, each auto-encoder is represented as φ_t , $t = 1, 2, \dots, N$, where the calculation of each auto-encoder is defined as:

$$\begin{cases} h_F = h_c + h_{\text{fusion}} \\ \Delta_{Z_t} = \varphi_t(h_F), & t = 1 \\ \Delta_{Z_t} = \varphi_t(h_{\text{fusion}} + \Delta_{Z_{t-1}}), & t > 1 \end{cases} \quad (3)$$

where Δ_{Z_t} represents the output of the t th auto-encoder, as shown in Fig. 6. The predicted multimodal embeddings of the missing visual modality based on specific and invariant features of existing available modalities can be expressed as $h' = \Delta_{Z_N}$.

3.4 Loss functions

During SWG-MMIN model training, the emotion recognition classification loss L_{cls} is used to supervise the training of emotion category targets. L_{cls} adopts cross-entropy loss to measure the difference between the model output Y and the target emotion category \hat{Y} . The loss can be expressed as:

$$L_{\text{cls}} = \text{CrossEntropy}(Y, \hat{Y}) \quad (4)$$

In addition, the performance of the SWG-MMIN model can be further improved by combining imagination loss L_{imagine} and invariant loss $L_{\text{invariance}}$. The imagination loss L_{imagine} is employed to minimize the discrepancy between the output h' of the imagination module and the ground-truth representations h'' . And the invariant loss $L_{\text{invariance}}$ aims to ensure the invariant feature H' of the full modality prediction as close as possible to the invariant feature H of the target modality continuously. In this paper, the Root Mean Square Error is used as a metric to guarantee the accuracy of the model for emotion recognition, where i represents the i th sample of video segments.

$$L_{\text{imagine}} = \text{RMSE}(h'_i, h''_i) \quad (5)$$

$$L_{\text{invariance}} = \text{RMSE}(H_i, H'_i) \quad (6)$$

The total loss function of the model is the sum of the classification loss, imagination loss, and invariance loss, shown as in formula (7). λ_1 and λ_2 are the weight factors. By minimizing the total loss function, the model can implement the optimization to the emotion recognition accuracy, imagination module generation ability and modality invariance. The total loss function is expressed as:

$$L = L_{\text{cls}} + \lambda_1 L_{\text{imagine}} + \lambda_2 L_{\text{invariance}} \quad (7)$$

4 Experimental results

4.1 Dataset

In this paper, the proposed SWG-MMIN model is verified on two datasets IEMOCAP [4] and MSP-IMPROV [5], respectively. Following the label division of MMIN Zhao, Li and Jin [37], both datasets are categorized into four emotion labels, that is, happy, angry, sad, and neutral. The split ratio of training set, validation set and test set is 8:1:1.

IEMOCAP [4] is a dataset containing recorded videos of dyadic conversations sessions. In each session, it contains scripted plays and spontaneous dialogues between a male speaker and a female speaker, with a total of 10 speakers in the database. This dataset offers approximately 12 h of audio-visual data, including video, audio and text. Following the emotional label processing method described in MMIN Zhao, Li and Jin [37], the dataset forms the four-class emotion recognition settings.

MSP-IMPROV [5] consists of 6 conversations by 12 English major students, a total of about 8438 utterance samples, and a total time of more than 9 h. The audio was recorded using two collar microphones at a sampling rate of 48 kHz and 32-bit PCM. The video was recorded by a digital camera with a resolution of 29.97 frames per second.

For all experiments on IEMOCAP, this paper uses weighted accuracy (WA) [3] and unweighted accuracy (UA) [10] as two evaluation metrics. For MSP-IMPROV experiments, due to the imbalance of emotion categories in this dataset, F1 scores are used as evaluation metrics.

4.2 Experimental settings

Similar to the raw features extraction method in MMIN Zhao, Li and Jin [37], the audio features x^a are 130-dimensional OpenSMILE [8] features configured as "IS13 ComParE". The visual features x^v are 342-dimensional "Denseface" features extracted by pretrained DenseNet [15]. The

Table 1 Multimodal emotion recognition results on IEMOCAP dataset under full-modality condition

Model	Train	Test	WA	UA
cLSTM-MMA [25]			0.7394	–
SSMM [18]			0.7560	0.7456
MMIN [37]			0.7651	0.7779
HFGCN [29]	{a,v,t}	{a,v,t}	0.7468	–
GraphSAGE [20]			0.6543	0.6640
CITN-DAF [9]			0.7750	–
SWG-MMIN			0.7772	0.7874

Best results are given in bold

Table 2 Multimodal emotion recognition results on the IEMOCAP dataset under missing-modality condition

Model	Metric	Testing conditions						
		{a}	{v}	{t}	{a, v}	{a, t}	{v, t}	average
MCTN [26]	WA	0.4975	0.4892	0.6242	0.5634	0.6834	0.6784	0.5894
	UA	0.5162	0.4573	0.6378	0.5584	0.6946	0.6834	0.5913
MMIN [37]	WA	0.5303	0.4864	0.6564	0.6395	0.7251	0.7082	0.6243
	UA	0.5440	0.4598	0.6691	0.6434	0.7435	0.7162	0.6293
MMIN [37]	WA	0.5658	0.5252	0.6657	0.6399	0.7294	0.7267	0.6410
	UA	0.5900	0.5160	0.6802	0.6543	0.7514	0.7361	0.6524
MRAN [22]	WA	0.5544	0.5323	0.6531	0.6470	0.7300	0.7211	0.6397
	UA	0.5701	0.4980	0.6642	0.6446	0.7458	0.7224	0.6408
IF-MMIN [40]	WA	0.5620	0.5197	0.6702	0.6533	0.7405	0.7268	0.6454
	UA	0.5813	0.5041	0.6820	0.6652	0.7544	0.7362	0.6490
SWG-MMIN	WA	0.5636	0.5351	0.6743	0.6521	0.7497	0.7354	0.6517
	UA	0.5855	0.5194	0.6845	0.6596	0.7623	0.7414	0.6587

Best results are given in bold

text features x^f are 1024-dimensional BERT word embeddings extracted by a pretrained Bert-large [7].

The output size of both the specificity encoder and the invariance encoder is 128. The Semantic-Wise Feature Guided Imagination module consists of 5 auto-encoders. For the SWG-MMIN model, this study uses the Adam optimizer [16] for training. The batch size is 64, the dropout rate is 0.5, the initial learning rate is 0.0002, and the learning rate is updated using LambdaLR [34].

The models are trained and evaluated on the IEMOCAP and MSP-IMPROV datasets with tenfold and 12-fold cross-validation, where each fold consisted of 60 epochs. Each model is run three times to alleviate the impact of parameter random initialization and verify the robustness of models. All models are implemented using the PyTorch and trained on a single RTX3090 GPU.

4.3 Comparison experiments

4.3.1 Full-modality comparison results

First, the SWG-MMIN model is compared with other advanced multimodal emotion recognition models under the full-modality condition on IEMOCAP dataset.

cLSTM-MMA [25]: It proposes a multimodal attention mechanism to model the correlation between three modalities, replacing concatenation.

SSMM [18]: It introduces a semi-supervised training strategy for discrete multimodal emotion recognition.

MMIN (Zhao, Li and Jin [37]): Its modality encoder network learns effective robust joint multimodal representations in both full and missing modalities for multimodal emotion recognition.

HFGCN [29]: It is a Hierarchical Fusion Graph Convolutional Network that learns more informative multimodal representations by considering modality dependencies.

GraphSAGE (J. [20]): It proposes a SER model for variable-length utterance modeling, aiming to maximize emotional information retention within the utterances.

CITN-DAF [9]: It enables parallel computation of modalities, and explores modal interactions using circulant matrices, enhancing feature integration with dimension-aware fusion.

The experimental results are shown in Table 1. The SWG-MMIN model proposed in this paper achieves better performance than other models. The recognition accuracy of SWG-MMIN are improved by 1.21% and 0.95% in terms of weighted accuracy (WA) [3] and unweighted accuracy (UA) [10] than MMIN (Zhao, Li and Jin [37]).

4.3.2 Uncertain missing-modality comparison results

The model is compared with other advanced models under different missing modality conditions on the IEMOCAP dataset.

MCTN [26]: It learns a joint representation through cyclic translation between missing and available modalities.

MMIN (Zhao, Li and Jin [37]): Its modality encoder network learns effective robust joint multimodal representations in both full and missing modalities for multimodal emotion recognition.

MMIN-Augmented baseline (Zhao, Li and Jin [37]): It pools the missing-modality training set and full-modality training set together to train the MMIN model.

Table 3 Multimodal emotion recognition results on the MSP-IMPROV dataset under missing-modality conditions

Model	Metric	Testing conditions						
		{a}	{v}	{t}	{a,v}	{a,t}	{v,t}	average
MCTN [26]	F1	0.3285	0.3810	0.5050	0.4683	0.5611	0.5886	0.4721
MMIN-Aug [37]	F1	0.4278	0.4185	0.5544	0.5396	0.6038	0.6295	0.5455
MMIN [37]	F1	0.4647	0.4471	0.5573	0.5740	0.6188	0.6411	0.5649
SWG-MMIN	F1	0.4664	0.4651	0.5808	0.5780	0.6205	0.6480	0.5722

Best results are given in bold

MRAN [22]: It proposes the Multimodal Embedding and Missing Index Embedding to guide the reconstruction of missing modalities features.

IF-MMIN [40]: It uses an invariant feature learning strategy for a missing modality imagination network.

The experimental results are shown in Table 2.

The SWG-MMIN model achieves the-state-of-the-art results on {v},{t},{v,t},{a,t} and the average of six missing modalities. Under the conditions {a} and {a, v}, SWG-MMIN is comparable to the best baseline. These results show that the SWG-MMIN model has excellent performance and robustness in performing emotion recognition tasks. From Table 2, it can be seen that the {v} modality obtains greater performance improvement than other missing modalities. It can be attributed to the CMFE module, which especially enhances the visual modality feature. In contrast, the performance under {a} and {a, v} conditions are comparable to the baseline, the possible reason why the audio modality lacks semantic information compared to the other two modalities. However, the proposed model can still learn robust joint multimodal representations and obtaining

Semantic-Wise Features based on enhanced modality features, resulting in robust performance under different missing conditions.

Next, the comparative experiment is conducted on MSP-IMPROV dataset, and the experimental results are shown in Table 3. The experimental results show that the proposed SWG-MMIN model achieves the-state-of-the-art F1 score under all missing-modality conditions. Even for weak modality {a}, there is also a certain improvement compared to the baseline model.

In summary, the proposed SWG-MMIN model shows excellent emotion recognition ability on two benchmark datasets, which verifies the robustness and generalization of the SWG-MMIN model.

4.3.3 Ablation study

In this section, ablation experiments are carried out on the SWG-MMIN model. To analyze the significance of each component in the SWG-MMIN model, we add an invariant feature semantic subspace based on CMD distance to the

Table 4 Results of ablation experiments of each module on IEMOCAP

CMFE	SWF	SWG I	Metric	Testing conditions							
				{a}	{v}	{t}	{a, v}	{a, t}	{v, t}	average	
√			WA	0.5529	0.5224	0.6638	0.6407	0.7343	0.7154	0.6382	
			UA	0.5735	0.5076	0.6771	0.6552	0.7515	0.7285	0.6489	
			WA	0.5565	0.5265	0.6663	0.6433	0.7376	0.7185	0.6414	
			UA	0.5780	0.5129	0.6817	0.6556	0.7529	0.7331	0.6523	
√	√		WA	0.5572	0.5289	0.6683	0.6431	0.7369	0.7264	0.6434	
			UA	0.5807	0.5139	0.6833	0.6560	0.7518	0.7379	0.6539	
	√		√	WA	0.5607	0.5310	0.6706	0.6458	0.7381	0.7302	0.6460
				UA	0.5829	0.5160	0.6841	0.6569	0.7552	0.7402	0.6558
√	√	√	WA	0.5628	0.5298	0.6728	0.6459	0.7431	0.7315	0.6476	
			UA	0.5816	0.5181	0.6843	0.6593	0.7571	0.7398	0.6567	
	√	√	WA	0.5636	0.5351	0.6743	0.6521	0.7497	0.7354	0.6517	
			UA	0.5855	0.5194	0.6845	0.6596	0.7623	0.7414	0.6587	

Best results are given in bold

Table 5 Results of the downsampled features of the CMFE module on IEMOCAP. Baseline+CMFE (x_1^m): add the CMFE module with a downsampled feature on the baseline; Baseline+CMFE (x_1^m, x_2^m): add the CMFE module with two downsampled features on the baseline; Baseline+CMFE (x_1^m, x_2^m, x_3^m): add the CMFE module with three

downsampled features on the baseline; Baseline+CMFE (x^m, x^m, x^m): add the CMFE module with three raws features on the baseline; SWG-MMIN- CMFE (x^m, x^m, x^m): the CMFE module with three raws features on the SWG-MMIN

Model	Metric	Testing conditions						average
		{a}	{v}	{t}	{a, v}	{a, t}	{v, t}	
Baseline	WA	0.5529	0.5224	0.6638	0.6407	0.7343	0.7154	0.6382
	UA	0.5735	0.5076	0.6771	0.6552	0.7515	0.7285	0.6489
Baseline+CMFE (x_1^m)	WA	0.5526	0.5228	0.6659	0.6398	0.7363	0.7179	0.6392
	UA	0.5729	0.5095	0.6798	0.6526	0.7519	0.7308	0.6496
Baseline+CMFE (x_1^m, x_2^m)	WA	0.5549	0.5245	0.6661	0.6411	0.7369	0.7182	0.6403
	UA	0.5748	0.5101	0.6811	0.6542	0.7521	0.7319	0.6507
Baseline+CMFE (x_1^m, x_2^m, x_3^m)	WA	0.5565	0.5265	0.6663	0.6433	0.7376	0.7185	0.6414
	UA	0.5780	0.5129	0.6817	0.6556	0.7529	0.7331	0.6523
Baseline+CMFE (x^m, x^m, x^m)	WA	0.5553	0.5226	0.6360	0.6424	0.7096	0.7147	0.6301
	UA	0.5748	0.5107	0.6497	0.6539	0.7129	0.7186	0.6368
SWG-MMIN-CMFE (x^m, x^m, x^m)	WA	0.5605	0.5332	0.6582	0.6458	0.7168	0.7305	0.6408
	UA	0.5763	0.5148	0.6621	0.6554	0.7286	0.7377	0.6458
SWG-MMIN	WA	0.5636	0.5351	0.6743	0.6521	0.7497	0.7354	0.6517
	UA	0.5855	0.5194	0.6845	0.6596	0.7623	0.7414	0.6587

Best results are given in bold

MMIN (Zhao, Li and Jin, 2021) as the baseline model. Then, Comprehensive Modality Feature Enrichment (CMFE) module, Semantic Efficient Fusion (SWF) module, and Semantic Wise Feature Guided Imaging (SWGFI) module are gradually applied to the model to verify the impact of each component on model performance. The experimental results are shown in Table 4.

Comprehensive Modality Feature Enrichment (CMFE) module. By adding the CMFE module to the baseline model, it can be observed that the model performance has improvement compared to the baseline model under various missing conditions. This indicates that the CMFE module has a significant effect on emotion recognition tasks. The effect proves the excellent performance and robustness of the CMFE module.

Semantic-Wise Fusion (SWF) module. Table 4 shows that by adding the SWF module, the performance of the model has been improved under most conditions. The effect under the condition of weak modality {a} and weak modality combination {a,v} is also improved to a certain extent. It proves that the SWF module plays an indispensable role in the SWG-MMIN model performing emotion recognition tasks.

4.4 Semantic-Wise Feature Guided Imagination (SWGFI) module

The construction of this module needs to rely on the adaptive fusion of modality-invariant and modality-specific features by the SWF module, so the ablation experiment is performed on the combination of SWF and SWGI modules. As shown in Table 4, on the basis of adding the SWF Module to the baseline model, the adaptive fused feature is fed into the SWGI module for generating imaginative representations. The performance of the model has been improved to a certain degree under different missing modality conditions. Among them, the improvement is most obvious under the condition of strong mode combination {a, t}. The effectiveness of the Semantic-Wise Feature Guided Imagination module is proved.

The above ablation experiments prove the effectiveness of each module in the SWG-MMIN model, as well as the complementary dependency relationships between the modules. This indicates that each component in the model plays a significant role.

Table 6 Results of different fusion methods on IEMOCAP dataset under full-modality condition. SWG-MMIN(MHSA): MHSA(Multi-Head Self-Attention) fusion method; SWG-MMIN(LAFF): LAFF(Lightweight Attentional Feature Fusion) block

Model	Train	Test	WA	UA
SWG-MMIN(MHSA)			0.7628	0.7744
SWG-MMIN(LAFF)	{a,v,t}	{a,v,t}	0.7605	0.7725
SWG-MMIN(SWF)			0.7772	0.7874

Best results are given in bold

4.4.1 Analysis of CMFE module

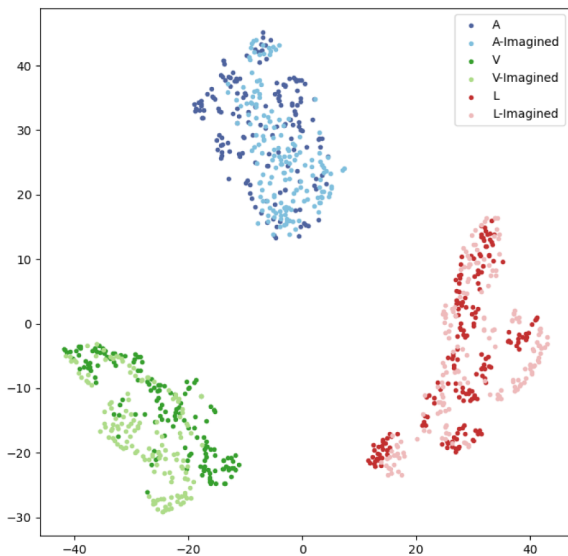
This section investigates the impact of different scale downsampled features in the CMFE module on model performance. On the basis of the raw features, this module adds

three downsampled features of different scales and fuses them to make up for the information loss of the raw features during the dimensionality reduction process. As shown in Table 5, as the number of downsampled features increases, the performance of the model also improves accordingly. However, due to computational considerations, this paper finally determines the number of downsampled features to 3 to achieve a balance between parameters and performance. In addition, this paper also explores the effectiveness of downsampled features at different scales in the CMFE module by replacing the downsampled features with scale-invariant raw features. The results show that model performance is adversely affected after replacing downsampled features, indicating that downsampled features at different scales can effectively compensate for information loss. While simply repeating features has a negative impact on performance.

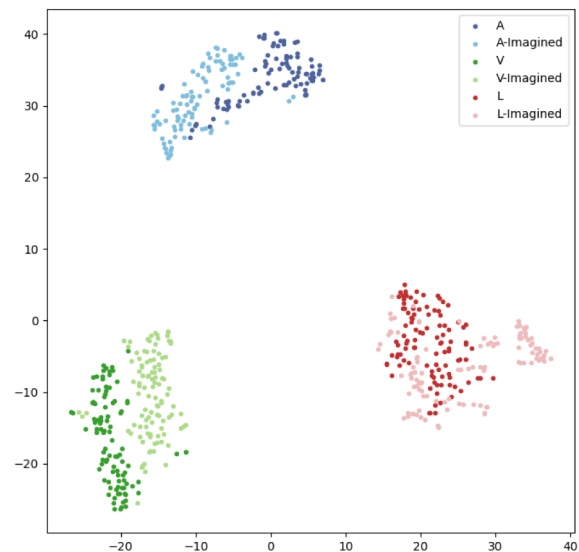
Table 7 Results of different fusion methods on IEMOCAP dataset under missing-modality conditions

Model	Metric	Testing conditions						average
		{a}	{v}	{t}	{a,v}	{a,t}	{v,t}	
SWG-MMIN (MHSA)	WA	0.5606	0.5227	0.6716	0.6365	0.7490	0.7269	0.6446
	UA	0.5836	0.5189	0.6843	0.6569	0.7601	0.7366	0.6567
SWG-MMIN (LAFF)	WA	0.5593	0.5385	0.6744	0.6396	0.7419	0.7365	0.6484
	UA	0.5832	0.5145	0.6888	0.6536	0.7555	0.7421	0.6563
SWG-MMIN (SWF)	WA	0.5636	0.5351	0.6743	0.6521	0.7497	0.7354	0.6517
	UA	0.5855	0.5194	0.6845	0.6596	0.7623	0.7414	0.6587

Best results are given in bold



(a) IEMOCAP



(b) MSP-IMPROV

Fig. 7 Visualization analysis of the ground-truth and SWG-MMIN imagined multimodal embeddings

Furthermore, it proves the important role of the CMFE module in enhancing model performance, which is not solely dependent on the SWF module. Downsampled features at different scales are also an indispensable part of the CMFE module. This highlights the necessity and effectiveness of different scales downsampled features in improving model performance.

4.4.2 Analysis of SWF module

This section studies the impact of different fusion methods on the performance of the SWG-MMIN model. Comparative experiments with other fusion methods are conducted to verify the effectiveness of the SWF adaptive fusion module. As shown in Table 6, under full-modality condition, the SWF fusion module is replaced with the MHSA (Multi-Head Self-Attention) fusion method and LAFF (Lightweight Attentional Feature Fusion)(F. [13]) block, respectively. It can be observed that the proposed SWG-MMIN (SWF) in this paper outperforms these two methods in both weighted accuracy (WA) and unweighted accuracy (UA). Thus, it demonstrates the effectiveness of the SWF fusion module proposed in this paper in improving model performance.

This section further conducts comparative experiments under different missing-modality conditions to more comprehensively verify the effectiveness of different fusion methods. The experimental results are shown in Table 7. It can be observed that the SWG-MMIN (SWF) proposed in this paper is better than other models in the overall average performance. And it also achieves the best performance under most missing-modality test conditions. Under the conditions $\{v\}$, $\{t\}$ and $\{v, t\}$, SWG-MMIN (SWF) is slightly weaker than SWG-MMIN (LAFF). The reason may be that the LAFF fusion module is designed for text-to-video retrieval tasks, performing better in text and video feature fusion but relatively worse in audio feature fusion. Therefore, the applicability of LAFF in this task is relatively low. The SWG-MMIN model proposed in this paper has achieved the best overall results. And the fusion module requires the adaptive fusion of six features, which is unsuitable for complex fusion networks. The SWF module is concise and performs well for this task, achieving an effective balance of accuracy and resources.

4.5 Visualization analysis

The t-SNE algorithm is used to randomly select sentences from the test sets of IEMOCAP and MSP-IMPROV. The aim is to visualize the ground-truth multimodal embeddings and SWG-MMIN-imagined multimodal embeddings in a two-dimensional plane under different missing modality conditions. As shown in Fig. 7, where A represents

the ground-truth multimodal embeddings of the audio modality, and A -Imagined represents SWG-MMIN imagined multimodal embeddings of the audio modality based on the visual and text modalities. It can be seen that the ground-truth multimodal embeddings of the three modalities are very similar to the SWG-MMIN imagined multimodal embeddings. And it demonstrates the effectiveness of the SWG-MMIN model on imagining the missing modalities based on the available modalities.

5 Conclusion

This study proposes a Semantic-Wise Guidance for the Missing Modality Imagination Network (SWG-MMIN). The SWG-MMIN model alleviates the modality gap by introducing the Comprehensive Modality Feature Enrichment module, Semantic-Wise Fusion module, and Semantic-Wise Feature Guided Imagination module. These modules fully utilize the semantic information of modality-invariant features and specific features to alleviate the heterogeneous modality gap and improve the robustness of joint multimodal representations. Experiments on the benchmark datasets IEMOCAP and MSP-IMPROV demonstrate the effectiveness and robustness of the SWG-MMIN model. It outperforms other baseline models in scenarios under full modalities and missing modalities. In future work, we will further explore methods to improve robust joint multimodal representation based on the fusion of modality-specific features and invariant feature.

Acknowledgements This work is supported partially by the project of Changchun Bureau of Science and Technology under Grant 21ZY31, the project of Jilin Provincial Science and Technology Department under Grant 20220201140GX, the project of Jilin Province Development and Reform Commission under Grant 2022C047-5. The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

Author contributions This study was completed by the co-authors. Shuhua Liu conceived the research. The major experiments and analyses were undertaken by Yixuan Wang and Kehan Wang. Binshuai Li drafted the related work. Shihao Yang was responsible for data processing and drawing figures. Fengqin Yang edited and reviewed the paper. All authors have read and approved the final manuscript.

Funding This work is supported by the National Natural Science Foundation of China under the Grant 62277009.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

References

1. Aguilar, G., Rozgic, V., Wang, W., and Wang, C.: Multimodal and multi-view models for emotion recognition. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 991–1002. <https://doi.org/10.18653/v1/P19-1095> 2019
2. Araño, K.A., Orsenigo, C., Soto, M., Vercellis, C.: Multimodal sentiment and emotion recognition in hyperbolic space. *Expert Syst. Appl.* **184**, 115507 (2021). <https://doi.org/10.1016/j.eswa.2021.115507>
3. Baidari, I., Honnikoll, N.: Accuracy weighted diversity-based online boosting. *Expert Syst. Appl.* **160**, 113723 (2020). <https://doi.org/10.1016/j.eswa.2020.113723>
4. 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.* **42**(4), 335–359 (2008). <https://doi.org/10.1007/s10579-008-9076-6>
5. Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., Provost, E.M.: MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception. *IEEE Trans. Affect. Comput. Syst.* **8**(1), 67–80 (2017). <https://doi.org/10.1109/TAFFC.2016.2515617>
6. Cai, L., Wang, Z., Gao, H., Shen, D., and Ji, S.: Deep adversarial learning for multi-modality missing data completion. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1158–1166. <https://doi.org/10.1145/3219819.3219963> 2018
7. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423> 2019
8. Eyben, F., Wöllmer, M., and Schuller, B.: Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462. <https://doi.org/10.1145/1873951.1874246> 2010
9. Gong, P., Liu, J., Zhang, X., Li, X., and Yu, Z.: Circulant-interactive transformer with dimension-aware fusion for multimodal sentiment analysis. 189. 2023
10. Gupta, S., Fahad, Md.S., Deepak, A.: Pitch-synchronous single frequency filtering spectrogram for speech emotion recognition. *Multimed. Tools Appl.* **79**(31–32), 23347–23365 (2020). <https://doi.org/10.1007/s11042-020-09068-1>
11. Han, J., Zhang, Z., Ren, Z., & Schuller, B.: Implicit fusion by joint audiovisual training for emotion recognition in mono modality. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5861–5865. <https://doi.org/10.1109/ICASSP.2019.8682773> 2019
12. Hazarika, D., Zimmermann, R., and Poria, S.: MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. *Proceedings of the 28th ACM International Conference on Multimedia*, 1122–1131. <https://doi.org/10.1145/3394171.3413678> 2020
13. Hu, F., Chen, A., Wang, Z., Zhou, F., Dong, J., & Li, X.: Light-weight attentional feature fusion: a new baseline for text-to-video retrieval (arXiv:2112.01832). arXiv. <http://arxiv.org/abs/2112.01832> 2022
14. Hu, J., Shen, L., and Sun, G.: Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141 <https://doi.org/10.1109/CVPR.2018.00745> 2018
15. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q.: Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269. <https://doi.org/10.1109/CVPR.2017.243> 2017
16. Kingma, D. P., and Ba, J.: Adam: a method for stochastic optimization. <https://doi.org/10.48550/ARXIV.1412.6980> 2014
17. Lian, Z., Chen, L., Sun, L., Liu, B., and Tao, J.: GCNet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–14. <https://doi.org/10.1109/TPAMI.2023.3234553> 2023
18. Liang, J., Li, R., and Jin, Q.: Semi-supervised multi-modal emotion recognition with cross-modal distribution matching. *Proceedings of the 28th ACM International Conference on Multimedia*, 2852–2861. <https://doi.org/10.1145/3394171.3413579> 2020
19. Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S.: Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944. <https://doi.org/10.1109/CVPR.2017.106> 2017
20. Liu, J., Wang, H., Sun, M., Wei, Y.: Graph based emotion recognition with attention pooling for variable-length utterances. *Neurocomputing* **496**, 46–55 (2022). <https://doi.org/10.1016/j.neucom.2022.05.007>
21. Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Bagher Zadeh, A., & Morency, L.-P.: Efficient low-rank multimodal fusion with modality-specific factors. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2247–2256. <https://doi.org/10.18653/v1/P18-1209> 2018
22. Luo, W., Xu, M., & Lai, H.: Multimodal reconstruct and align net for missing modality problem in sentiment analysis. In D.-T. Dang-Nguyen, C. Gurrin, M. Larson, A. F. Smeaton, S. Rudinac, M.-S. Dao, C. Trattner, & P. Chen (Eds.), *MultiMedia Modeling* (Vol. 13834, pp. 411–422). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-27818-1_34 2023
23. Lv, F., Chen, X., Huang, Y., Duan, L., and Lin, G.: Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2554–2562. <https://doi.org/10.1109/CVPR46437.2021.00258> 2021
24. Mun, J., Cho, M., and Han, B.: Text-guided attention model for image captioning. <https://doi.org/10.48550/ARXIV.1612.03557> 2016
25. Pan, Z., Luo, Z., Yang, J., and Li, H.: Multi-modal attention for speech emotion recognition. <https://doi.org/10.48550/ARXIV.2009.04107> 2020
26. Pham, H., Liang, P. P., Manzini, T., Morency, L.-P., and Póczos, B.: found in translation: learning robust joint representations by cyclic translations between modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 6892–6899. <https://doi.org/10.1609/aaai.v33i01.330168922019> 2019
27. Poklukar, P., Vasco, M., Yin, H., Melo, F. S., Paiva, A., and Kragic, D.: Geometric multimodal contrastive representation learning. <https://doi.org/10.48550/ARXIV.2202.03390> 2022
28. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. *Inform. Fusion* **37**, 98–125 (2017). <https://doi.org/10.1016/j.inffus.2017.02.003>
29. Tang, S., Luo, Z., Nan, G., Baba, J., Yoshikawa, Y., and Ishiguro, H.: Fusion with hierarchical graphs for multimodal emotion recognition. *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1288–1296. <https://doi.org/10.23919/APSIPAASC55919.2022.9979932> 2022
30. Tran, L., Liu, X., Zhou, J., and Jin, R.: Missing Modalities Imputation via Cascaded Residual Autoencoder. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4971–4980. <https://doi.org/10.1109/CVPR.2017.528> 2017

31. Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569. <https://doi.org/10.18653/v1/P19-1656> 2019
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: *Attention is all you need*. <https://doi.org/10.48550/ARXIV.1706.03762> 2017
33. Wang, Z., Wan, Z., and Wan, X.: TransModality: An end2end fusion method with transformer for multimodal sentiment analysis. *Proceedings of The Web Conference 2020*, 2514–2520. <https://doi.org/10.1145/3366423.3380000> 2020
34. Wu, N., Green, B., Ben, X., and O'Banion, S.: *Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case*. <https://doi.org/10.48550/ARXIV.2001.08317> 2020
35. Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P.: Tensor fusion network for multimodal sentiment analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114. <https://doi.org/10.18653/v1/D17-1115> 2017
36. Zhang, M., Mosbach, M., Adelani, D., Hedderich, M., & Klakow, D.: MCSE: Multimodal contrastive learning of sentence embeddings. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5959–5969. <https://doi.org/10.18653/v1/2022.naacl-main.436> 2022
37. Zhao, J., Li, R., and Jin, Q.: Missing modality imagination network for emotion recognition with uncertain missing modalities. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2608–2618. <https://doi.org/10.18653/v1/2021.acl-long.203> 2021
38. Zhao, J., Li, R., Jin, Q., Wang, X., and Li, H.: *MEmoBERT: Pre-training model with prompt-based learning for multimodal emotion recognition*. <https://doi.org/10.48550/ARXIV.2111.00865> 2021
39. Zhu, L., Zhu, Z., Zhang, C., Xu, Y., Kong, X.: Multimodal sentiment analysis based on fusion methods: a survey. *Information Fusion* **95**, 306–325 (2023). <https://doi.org/10.1016/j.inffus.2023.02.028>
40. Zuo, H., Liu, R., Zhao, J., Gao, G., & Li, H.: Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. <https://doi.org/10.1109/ICASSP49357.2023.10095836> 2023

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.