Expert Systems With Applications 249 (2024) 123655

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

MIT-FRNet: Modality-invariant temporal representation learning-based feature reconstruction network for missing modalities

Jiayao Li [a], Saihua Cai [b], Li Li [c], Ruizhi Sun [a,d,*], Gang Yuan [a], Rui Zhu [a]

[a] College of Information and Electrical Engineering, China Agricultural University, Beijing 110000, China
[b] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China
[c] Computer School, Beijing Information Science and Technology University, Beijing, 100101, China
[d] Scientific Research Base for Integrated Technologies of Precision Agriculture (Animal Husbandry), The Ministry of Agriculture, Beijing 110000, China

ARTICLE INFO

Keywords:
Feature reconstruction
Missing modalities
Modality-invariant representation
Data fusion

ABSTRACT

The investigation of missing modalities aims to extract valuable feature information from missing multi-modal data, it is a focal point in multi-modal learning. Existing missing modalities processing methods primarily focused on multi-modal fusion schemes to achieve optimal performance, but they faced the following two key challenges: (1) how to improve the robustness of incomplete multi-modal sequence representation, (2) how to effectively learn modality-invariant representations to mitigate heterogeneity between modalities. In this paper, we propose a modality-invariant temporal representation learning-based feature reconstruction network called MIT-FRNet for the missing modalities to tackle these challenges. In the MIT-FRNet, we first extract the latent features for each modality considering intra-modality and inter-modality, and then introduce an encoder-decoder framework to address the first challenge, it reconstructs missing element features via taking the incomplete modal sequences as input and implementing the inter-modal and cross-modal attention mechanisms for feature extraction. And then, through treating each timestamp as a single Gaussian distribution, we design a fine-grained similarity constraint based on distribution-level modality-invariant representations to learn effective modality-invariant representations, thereby addressing the second challenge. Finally, the efficiency of proposed model is validated through the classification results after multi-modal fusion that involves using a gate encoder to pass it and followed by a vector fusion to fuse it. Extensive experiments on public benchmark datasets demonstrate that the proposed MIT-FRNet method achieves promising results under varying missing rates while exhibiting good convergence.

1. Introduction

In recent years, with the diversification of information, multi-modal data (including language, acoustics, visual, etc.) has become more and more common in real life. Every modality contributes information (modality-specific features and modality-invariant features) in the prediction. Combining different modalities allows for the learning of complementary features and obtaining better joint multi-modal representations, resulting in the improvement performance in the prediction tasks. Compared with unimodal data, multi-modal data is prone to missing data or features during acquiring, transmitting and storing. As illustrated in Fig. 1, the failure of sensors leads to the loss of data during acquisition, which causes not collecting or adding noise of acoustic data, the missing of visual data, and making text data becoming garbled codes or missing. In past researches, the incomplete modal sequences negatively impact the performance of multi-modal model (Wei, Luo, Ma, Ren, & Luo, 2023; Zhou, Ruan, & Hu, 2022), the uncertain faults lead to missing and incomplete of modal data, and the loss of modal data leads to a decrease of modal features, thereby resulting in multi-modal feature sparsity. In addition, multi-modal data are usually missing at random. Although there are some researches (Matsuura, Saito, Ushiku, & Harada, 2018; Wang, Ding, Tao, Gao, & Fu, 2021; Wei et al., 2023) for randomly missing multi-modal data, but they have significant shortcomings. Thus, the following challenges still exist in multi-modal learning of missing modalities: (1) The feature information of sparse sequences of missing modalities is failing to be fully captured; (2) And complementary

* Corresponding author.
E-mail addresses: lijiayao@cau.edu.cn (J. Li), caisaih@ujs.edu.cn (S. Cai), lili211213@bistu.edu.cn (L. Li), sunruizhi@cau.edu.cn (R. Sun), yuangang@cau.edu.cn (G. Yuan), ruiz@cau.edu.cn (R. Zhu).

https://doi.org/10.1016/j.eswa.2024.123655
Received 10 October 2023; Received in revised form 29 February 2024; Accepted 9 March 2024
Available online 11 March 2024

information in missing modal data cannot be adequately learnt in multi-modal learning. Thus, to simultaneously process the randomly missing multi-modal data, it is necessary to map different modalities into a common space. Therefore, enhancing and integrating data from multiple modalities shows paramount importance (Hazarika, Zimmermann, & Poria, 2020; Liu, Zhao, Wei, Zheng, & Yang, 2019; Sun, Liu, Chen, & Lin, 2023).

Semantic sparsity in incomplete modal sequences is an important cause of missing random modalities, it poses a challenge in extracting robust modal representations. Previous researches on multi-modal have explored various methods to deal with randomly missing data. For example, GAN (generative adversarial network)-based methods (Mirza & Osindero, 2014; Shang et al., 2017; Zhang, Shen, Zhang, & Wang, 2021) capture inter-modal relationships through generation network to estimate missing modalities and reduce the modal gaps (Qian & Wang, 2023); Correlation-based methods (Ma, Huang, & Zhang, 2021; Matsuura et al., 2018; Mittal, Bhattacharya, Chandra, Bera, & Manocha, 2020) exploit correlations between non-missing and missing modalities to estimate the latter; Cyclic consistency-based methods (Pham, Liang, Manzini, Morency, & Póczos, 2019; Wang et al., 2021; Zhao, Li, & Jin, 2021) learn joint representations by modelling relationships between modalities and align semantic information across multiple modalities (Li et al., 2022; Zhang, Wang, & Liu, 2023). Currently, most generative methods used to deal with missing modalities still rely on encoder-based reconstruction (Baldi, 2012; Cai, Wang, Gao, Shen, & Ji, 2018; Kingma & Welling, 2013; Liu, Zhou, Chu, Sun, & Meng, 2024; Tran, Liu, Zhou, & Jin, 2017; Yuan, Li, Xu, & Yu, 2021; Zeng, Zhou, & Liu, 2022; Zhang et al., 2020). Existing methods focus on extracting the relationship between missing modalities and non-missing modalities, while ignoring the shared features that exist between them. In contrast to these methods, our research employs an encoder-decoder structure for multi-modal representation learning, which generates hidden representations and captures shared semantics from source and target sequences (Guo, Wang, & Wang, 2019).

On the other hand, current researches on random missing modalities fail to consider the significant heterogeneity among different modalities, which greatly impacts the performance of fusion (Bai, Chen, Zhou, Yi, & Chien, 2021; Chauhan, Akhtar, Ekbal, & Bhattacharyya, 2019; Xu, Meng, Qiu, Yu, & Wu, 2019). To address this issue, researchers typically learn modality-invariant representations from relevant modal features (Hazarika et al., 2020; Leidal, Harwath, & Glass, 2017; Liu et al., 2019; Peng, Zhang, & Huang, 2019) before fusion. As is shown in Fig. 2, three modalities including language, acoustic and visual present heterogeneity. The modality-specific representations make each modality showing specific features, which requires learning the modality-invariant representations to align semantic features and improve the fusion features, thereby mitigating the impact of multi-modal heterogeneity on fusion. In the existing methods, although measure methods based on cosine similarity or other specific similarities (Sun, Sarma, Sethares, & Liang, 2020; Thongtan & Phienthrakul, 2019) ensure that the features between modalities in the shared projection space are closer, they also suffer from information loss. The optimizing central moment discrepancy (CMD) (Zellinger et al., 2019) can avoid this problem by mitigating the distribution-level differences between modes, but it ignores the order information of modalities. To solve the loss of important specific information and consider sequence information, finer-grained similarity constraint is employed.

In summary, the existing missing modality processing methods still face the following two problems that need to be solved urgently: (1) When modalities are missing, the incomplete multi-modal sequences can only extract sparse features and thus weakening the fusion performance, and the missing of modalities can also decrease the robustness of multi-modal representation; (2) Due to the heterogeneity of modalities, the specific representations of modality interference the features in common space for fusion, which has a negative impact on the fusion performance for multi-modal.

For the existing approaches in multi-modal learning with missing modalities, multiple discriminators modelling complex correlations between different views based on cross-view networks are used to generate missing data (Zhang et al., 2020). Modal specificity is weakened by enhancing modal coherence (Hazarika et al., 2020) while attention-based acquisition of a robust representation of each element in a modal sequence improves the model prediction ability in unaligned modal sequences (Yuan et al., 2021). They lack exploration of multi-modal common features and have not considered finer-grained information within multi-modal contexts. To solve the above two limitations, we propose a modality-invariant temporal representation learning-based feature reconstruction network called MIT-FRNet for missing modalities. Firstly, we adopt intra-modal and inter-modal attention-based extractors to learn robust representations for each element in the modality sequences, and then propose a reconstruction module to generate the missing modality features, thereby improving the robustness of model for the random missing in modality sequences. Secondly, we treat each modality as a multivariate Gaussian distribution (considering each timestamp as a single Gaussian distribution) and use the Kullback-Leibler (KL) divergence to capture the implicit temporal distribution-level similarities, thereby learning modality-invariant representations. These strategies not only help for improving the robustness of model for the random missing in non-aligned modality sequences, but also reduce domain shifts between different modalities as well as extract effective sequential modality-invariant representations.

The major contributions of this paper are as follows:

1. We propose a transform-based feature reconstruction network to improve the model robustness against random missing in modal sequences. We employ extractors based on intra-modal and inter-modal attention mechanisms to acquire robust representations of each element within the modal sequence. Additionally, we introduce a reconstruction module for generating missing modal features, enabling us to learn semantic-level features corresponding to these missing elements.
2. We propose a modality-invariant temporal representation that mitigates the heterogeneity of multi-modal data and improves the multi-modal representation, where each timestamp feature of each modality is considered as a Gaussian distribution and the multivariate distribution KL divergence is utilized to constrain similarity at the

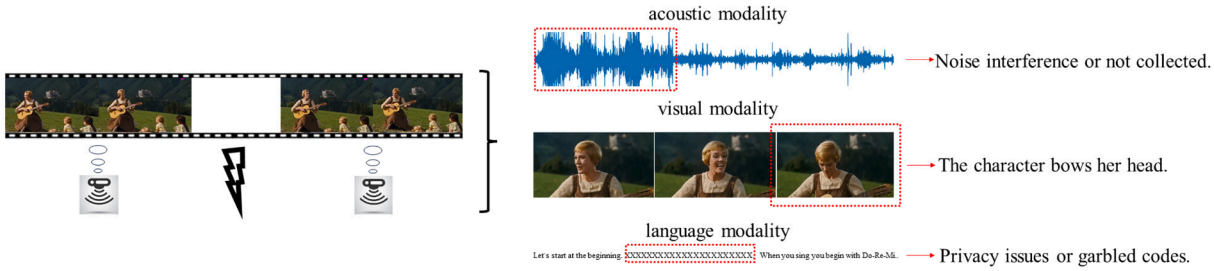

Fig. 1. A case of missing modalities.

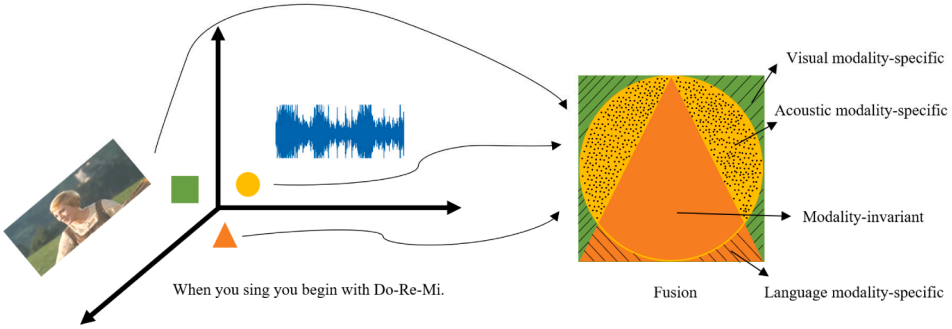

Fig. 2. Learning multi-modal representations through modality-invariant and -specific subspaces for fusion.

temporal level, thereby ensuring consistency across different modalities over time. Additionally, we leverage sequence information to extract similarity from multi-modal features and prevent loss of specific information during this process.

3. We conduct extensive experiments on two typically benchmark datasets to evaluate the proposed MIT-FRNet model. The experimental results indicate that compared with existing models, MIT-FRNet achieves better performances (including accuracy, F1 Score, MAE, correlation coefficient and running time) and has good convergence and stability. Code and data available at https://github.com/xiajili/MIT-FRNet.

The remaining parts of this paper are organized as follows. Section 2 introduces the related work. Section 3 describes the MIT-FRNet model, a modality-invariant temporal representation learning-based feature reconstruction network for missing modalities. Section 4 presents the experimental results and analysis. Section 5 concludes this paper and discusses the future work.

2. Related work

In this section, we introduce the multi-modal learning for missing modalities as well as the modality-invariant learning representation.

2.1. Multi-modal learning for missing modalities

The existing multi-modal learning for missing modalities has divided into four categorizations, including GAN (generative adversarial network)-based methods, correlation-based methods, cycle-based methods and encoder-based methods.

GAN-based methods: The GAN-based methods estimated the missing modalities via capturing the relationship between modalities using generative adversarial networks (Mirza & Osindero, 2014). For example, Shang et al. (2017) proposed a missing view imputation with GAN, namely VIGAN, it used the randomly sampled data of each view to identify domain-to-domain mapping through GAN, which were considered as imputation of missing data. Zhang et al. (2021) proposed a partial modal conditioned GANs for multi-modal multi-label learning with arbitrary modal-missing (called PMC-GAN), it combined all available modalities to generate high-quality missing modalities. Qian and Wang (2023) proposed a contrastive masked-attention model (COM) for incomplete multimodal learning. The reduction of modal gaps was achieved through GAN-based enhancement in cross-modal contrastive learning, while the interactions between modalities were captured via employing a masked-attention model. The GAN-based methods usually ignore the correlation between missing modalities and non-missing modalities, which will influence the generation of missing modalities.

Correlation-based methods: Correlation-based methods exploit correlations between non-missing and missing modalities to estimate missing modalities. In recent years, many scholars proposed correlation-based multi-modal learning methods for missing modalities based on the correlation between the missing modalities and non-missing modalities. For example, Mittal et al. (2020) proposed a multiplicative multimodal emotion recognition method called M3ER to judge whether the input modality was a missing modality, and the proxy feature vector for the missing modalities was generated for prediction; Ma et al. (2021) proposed a correlation loss based on Hirschfeld-GebeleinŔenyi (HGR) maximum correlation, it captured the common information to deal with missing modalities; In addition, a generalized Bayesian canonical correlation analysis with missing modalities called GBCCA-M2 (Matsuura et al., 2018) was proposed by Matsuura et al. in 2018, it included incomplete set of modalities in the likelihood function to learn the relationship between missing and non-missing modalities. However, correlation-based methods are limited in deal with multiple missing modalities at the same time.

Cycle-based methods: Cycle-based methods learn the joint representation by modelling the relationship between different modalities based on cycle-consistency. For example, based on the idea that cycle-consistency loss can preserve the maximum information of all modalities, Pham et al. (2019) learned robust representations through cyclic transformations from source to target modalities based on the; Zhao et al. (2021) applied cycle-consistency learning to the imputation of missing modalities; Wang et al. (2021) proposed a generative partial multi-view clustering model called GP-MVC with adaptive fusion and cycle consistency, it explicitly generated data with missing views based on the shared representations provided by other views. Li et al. (2022) proposed a deep multimodal adversarial cycle-consistent network (DMACCN) composed of the modality specific-encoder, the modality-common fusion network, the cycle-consistent modality-specific generator, and the modality-fusion discriminator. It can fully fuse complementary information of data and align semantics between modalities and capture clustering structures of instances by an adversarial cycle-consistent loss. Zhang et al. (2023) introduced the cycle consistency constraint into region-phrase pairs to strengthen correlated pairs and weaken unrelated pairs. The bidirectional association between image regions and text phrases was used in cycle pairing to alleviate matching ambiguity. Unfortunately, cycle-based methods only focus on the fixed missing modalities but fail to consider the case of random missing modalities.

Encoder-based methods: At present, most methods used to obtain features of missing modalities rely on encoder-based reconstruction. Vincent, Larochelle, Bengio, and Manzagol (2008) used autoencoder (AE) to extract features, it was robust to partial damage of input data based on representation learning. Baldi (2012) used AE to learn the latent representations, Kingma and Welling (2013) designed a variational encoder by inferring and learning features from raw samples. Tran et al. (2017) proposed a cascaded residual autoencoder (CRA) to estimate data with missing modalities, it combined a series of residual AEs into a cascaded architecture to learn the relationship between different modalities. Cai et al. (2018) formulated the missing modalities problem as a conditional image generation task, and then designed a 3D encoder-decoder network to capture the modality relationship as well as put the

available category information during training into the network to enhance the robustness of the model. Zhang et al. (2020) developed a cross-part multi-view network to model complex correlations between different views, where multiple discriminators were used to generate missing data. Yuan et al. (2021) used Transformer to extract intra-modal and inter-modal relations, and then designed a Transformer-based feature reconstruction network to reproduce the semantics of missing modalities. Zeng et al. (2022) proposed a tag-assisted transformer encoder (TATE) network for missing uncertain modalities. A new space projection pattern was adopted to align common vectors, and a Transformer encoder-decoder network was utilized to learn the missing modality features. Liu et al. (2024) proposed a modality translation-based MSA model (MTMSA). The missing joint features were encoded by the transformer encoder module, which make them approximate those of complete modalities. Although encoder-based methods ignore the shared features, but the missing modalities can be reconstructed based on the encoder, which is a commonly generative way to deal with the missing modalities.

In summary, these methods exploited the relationship between missing and non-missing modalities to generate missing modalities through co-learning or utilizing the relationships, but they failed to consider the shared semantic features between missing and non-missing modalities. Consequently, this study aims to semantically reconstruct missing modalities by developing an innovative encoder-decoder framework. In addition, it also aims to capture the relationships between missing and non-missing modalities as well as encodes on multiple random missing modalities. The comparison of different multi-modal learning models is shown as Table 1.

2.2. Modality-invariant learning representation

The modality-invariant representation extracts common features from different modalities. As the research progresses, the researches in this area can be categorized: feature-level similarity constraint-based methods and distribution level similarity constraint-based methods.

Feature-level similarity constraint-based methods: This category of methods learns modality-invariant representation using feature level-based similarity constraints, such as cosine similarity (Thongtan & Phienthrakul, 2019). While the semantic similarity measures between texts and words were calculated using a custom model (Araque, Zhu, & Iglesias, 2019). This numeric-level invariant constraint representation was unstable, which would lead to information loss.

Distribution-level similarity constraint-based methods: This category of methods treats the features of each modality as distributions and applies the constraints at the distribution level, where each modality is projected onto two different subspaces: one space learns the commonalities between modalities and the other space captures the distinct features of each modality (Hazarika et al., 2020). For example, Zellinger et al. (2019) maximized the similarity between domain-specific activation distributions to achieve potential feature representations of domain invariance. In addition, the weighted domain-invariant representation learning (WDIRL) framework (Peng et al., 2019) and adversarial learning-based method (Liu et al., 2019) were also proposed for modality invariant representation. A cross-modal adversarial network (CMAN) (Leidal et al., 2017) was proposed that combines cross-modal adversarial learning with modality invariant attention learning to improve semantic alignment and answer prediction accuracy by learning modality invariant features. However, due to different modalities exhibit varying features across timestamps, single-variable distribution-level feature constraints fail to capture their hidden representations at the temporal level.

In summary, for the former methods, the use of feature-level similarity constraints may result in the loss of important specific information; the latter methods neglect of sequence information and reduction model performance. Consequently, a multi-modal temporal modality-invariant representation learning network is proposed to consider each timestamp as a single distribution and to capture the implicit temporal distribution-level similarities for learning modality-invariant representations.

Overall, previous studies on missing modalities have failed to achieve enhanced learning of common semantic features between missing and non-missing modalities as well as consider modality-invariant representations at multi-modal time-level. To address these issues, we propose an encoder-decoder structure for reconstructing missing modalities to facilitate enhanced common feature learning; In addition, we design a multi-modal temporal modality-invariant representation in the network to alleviate multi-modal heterogeneity.

3. MIT-FRNet

The overview framework of proposed MIT-FRNet (Modality-Invariant Temporal representation learning-based Feature Reconstruction Network for Missing Modalities) is shown in Fig. 3, it is composed of the following four parts: (1) Multi-modal feature extraction, it is used to extract potential features from diverse modalities; (2) Missing modal reconstruction, it is used to reconstruct the absent modalities via transformer encoder; (3) Modality-invariant temporal representation learning, it is used to acquire modality-invariant representations at the temporal level with refiner-grained similarity constraints; (4) Multi-modal fusion, it is used to integrate and predict the fused features across different modalities.

In a video segment $s \in S$, $X = (X^l, X^a, X^v)$ represents the raw multi-

Table 1
The comparison of different multi-modal learning models.

Models	Categories	Considering modality-invariant	Prediction Accuracy	Missing rates
VIGAN (Shang et al., 2017)	GAN-based	×	Relatively high	–
PMC-GAN (Zhang et al., 2021)	GAN-based	×	Medium	{0.0,0.1,0.2,0.3,0.4,0.5}
COM (Qian et al., 2023)	GAN-based	×	Medium	{0.0,0.1,0.2,0.3,0.4,0.5}
M3ER (Mittal et al., 2020)	Correlation-based	×	Relatively high	–
HGR (Ma et al., 2021)	Correlation-based	×	Relatively high	{0.4,0.6,0.8}
GBCCA-M2 (Matsuura et al., 2018)	Correlation-based	×	Low	{elements}
MCTN (Pham et al., 2019)	Cycle-based	×	Medium	–
MMIN (Zhao et al., 2021)	Cycle-based	×	Medium	{0.1,0.3,0.5,0.7,0.9}
GP-MVC (Wang et al., 2021)	Cycle-based	×	Medium	–
DMACCN (Li et al., 2022)	Cycle-based	×	Medium	–
WSVG (Zhang et al., 2023)	Cycle-based	×	Medium	–
CRA (Tran et al., 2017)	Encoder-based	×	Medium	{0,0.1,0.2,0.3,0.4,0.5}
CPM-Nets (Zhang et al., 2020)	Encoder-based	×	Relatively High	{0,0.1,0.2,0.3,0.4,0.5}
TFR-Net (Tran et al., 2017)	Encoder-based	×	Relatively High	{0.0–1.0}
MISA (Hazarika et al., 2020)	Encoder-based	√	Relatively High	–
TATE (Zeng et al., 2022)	Encoder-based	×	Relatively High	{0,0.1,0.2,0.3,0.4,0.5}
MTMSA (Liu et al., 2024)	Encoder-based	×	Relatively High	{0,0.1,0.2,0.3,0.4,0.5}
MIT-FRNet (ours)	Encoder-based	√	High	{0.0–1.0}

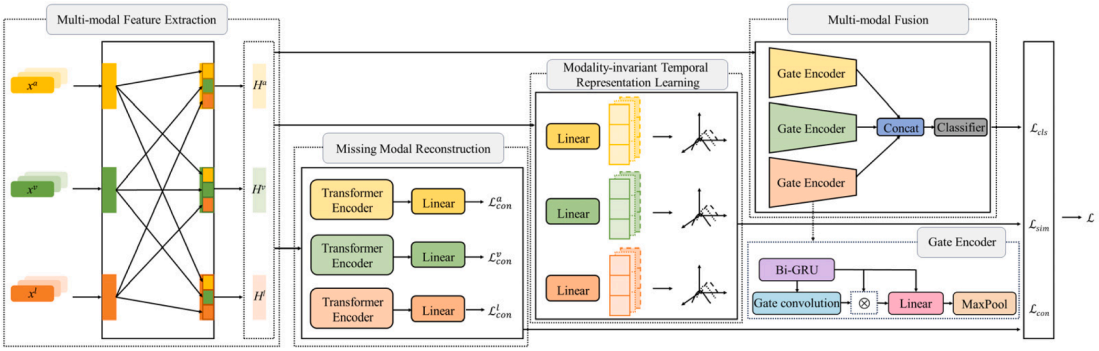

Fig. 3. The overall framework of MIT-FRNet model.

modal feature, X^l, X^a and X^v represent the features of language, acoustic and visual, respectively. $X_m^l, X_m^a and X_m^v$ represent the missing modality of language, acoustic and visual, where m means the missing modality. Assuming that the visual modality and language modality are absent, then the multi-modal representation can be denoted as $X_m^{'}$.

3.1. Multi-modal feature extraction

In the MIT-FRNet model, the first important task is to extract the multi-modal features. Unlike other methods only focus on inter-modal features to improve the accuracy but disregard intra-mode information, our approach effectively captures the important specific information between intra-modal and inter-modal, and also extracts the features between different modality pairs, thereby enhancing the prediction accuracy. Our model retains crucial details by dynamically adjusting attention scope and capturing sequence dependencies, thereby improving the representation ability and accuracy of the model. To effectively extract both inter-modal and intra-modal features, we propose a modal feature extraction sub-network, which is shown in Fig. 4.

In the multi-modal features extraction, firstly, a designed 1D-convolution is used to capture the sequence information at the temporal level. The nearest neighbor points of input modal sequence are captured by the convolution layer due to 1D-convolution is capable of analyzing the relationship between sequential adjacent feature components and considering global information. On the basis of 1D-convolution, we design the kernel and padding to ensure consistent sequence length and to obtain special adjacent information. With the special design, 1D-convolution perceives both global and local information at the temporal level. The convolution result is shown in formula (1).

$$C_l = Conv1d(X_m^l, kal^l), C_a = Conv1d(X_m^a, kal^a), C_v = Conv1d(X_m^v, kal^v) \quad (1)$$

In formula (1), kal^l, kal^a, kal^v mean the convolutional kernels of language, acoustic and visual respectively. We put each modality into the convolution to obtain feature information for the different modalities.

Intra-modal feature extraction. Secondly, the three intra-modal features are extracted respectively by intra-modal encoder. The intra-modal information is encoded via intra-modal encoder to capture the relationships of inter-sequence, where the position embedding module is utilized to improve the convolutional sequences and extract features across different modalities. The obtained result with intra-modality feature extraction is shown in formula (2).

$$\begin{aligned} X_{l\to l} &= f^{l,l}_{transformer}(C_l, C_l, C_l), X_{a\to a} = f^{a,a}_{transformer}(C_a, C_a, C_a), X_{v\to v} \\ &= f^{v,v}_{transformer}(C_v, C_v, C_v) \end{aligned} \quad (2)$$

In formula (2), $f(\cdot)$ means the transformer encoder of position embedding. $X_{l\to l}$, $X_{a\to a}$, $X_{v\to v}$ are the extracted features within the modalities of language, acoustic and visual respectively. Queries, keys, and values are the inputs of intra-encoder, and the sources are from the same modality. It ensures the features are extracted from intra-modal

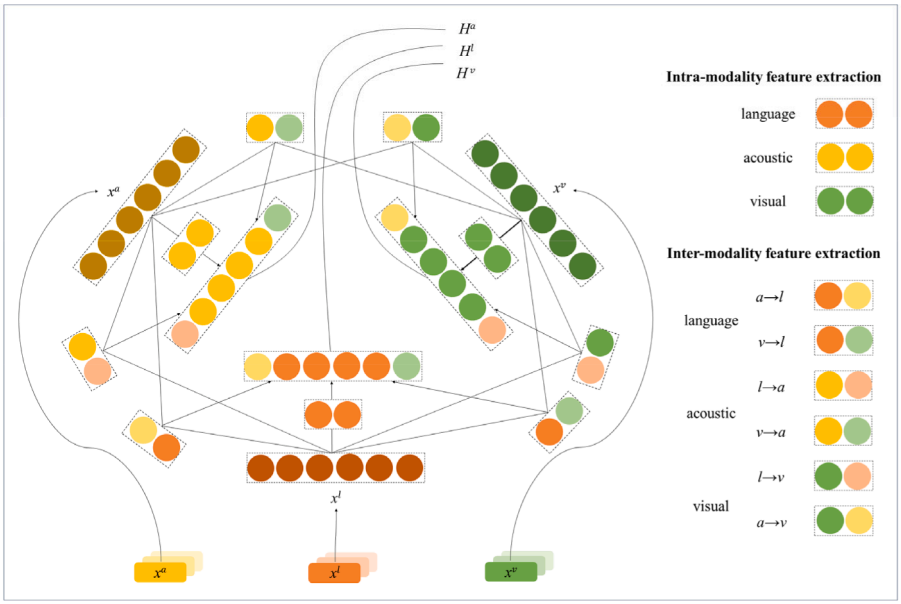

Fig. 4. An illustration of multi-modal feature extraction.

sequence.

Inter-modal feature extraction. Then, inter-modal features are obtained by tackling modality pairs through inter-modal encoder, and the modality pairs represent the correspondence between the modality and the other two modalities. The inter-modal features are obtained by different modalities with the use of inter-modal encoder to handle various modality pairs. To achieve this, we design specific modality pairs for language, acoustic and visual. The language modality pairs include the language-acoustic (a $\rightarrow l$) and language-visual ($v \rightarrow l$), the acoustic modality pairs include the acoustic-language ($l \rightarrow a$) and acoustic-visual ($v \rightarrow a$), and the visual modality pairs include visual-language ($l \rightarrow v$) and visual-acoustic ($a \rightarrow v$). In addition, we employ a transformer encoder to extract features between different modalities, where the encoder takes pairwise modality as its inputs. This approach facilitates interaction among diverse modal features and improves global semantic during information acquisition.

Specifically, the obtain of inter-modal features for language modality from the pairs of language-acoustic (a $\rightarrow l$) and language-visual ($v \rightarrow l$) modalities is shown in formula (3).

$$X_{a\rightarrow l} = f^{a,l}_{transformer}(C_l, C_a, C_a), X_{v\rightarrow l} = f^{v,l}_{transformer}(C_l, C_v, C_v) \tag{3}$$

The obtain of inter-modal features for acoustic modality from the pairs of acoustic-language ($l \rightarrow a$) and acoustic-visual ($v \rightarrow a$) modalities is shown in formula (4).

$$X_{l\rightarrow a} = f^{l,a}_{transformer}(C_a, C_l, C_l), X_{v\rightarrow a} = f^{v,a}_{transformer}(C_a, C_v, C_v) \tag{4}$$

Similarly, the obtain of inter-modal features for visual modality from the pairs of visual-language ($l \rightarrow a$) and visual-acoustic ($v \rightarrow a$) modalities is shown in formula (5).

$$X_{l\rightarrow v} = f^{l,v}_{transformer}(C_v, C_l, C_l), X_{a\rightarrow v} = f^{a,v}_{transformer}(C_v, C_a, C_a) \tag{5}$$

In formulas (3)–(5), $f(\cdot)$ means the encoder of position embedding. Queries, keys, and values are the inputs of inter-encoder, and the sources are from the different modalities. The features are extracted from inter-modal sequence common space. The source of queries is from a certain modality while the source of keys and values is from the other one. The inter-modal feature sequences are extracted via the effective representation taking advantage of the complementarity between modalities.

Finally, the modalities features are integrated including its intra-modal features and inter-modal features. All latent features obtained with all intra-modal and inter-modal encoders are concatenated as the enhanced sequence features output. The intra-modal and inter-modal features are integrated with formula (6).

$$\begin{aligned} H_l &= Concat([X_{l\rightarrow l}; X_{a\rightarrow l}; X_{v\rightarrow l}]), H_a = Concat([X_{a\rightarrow a}; X_{l\rightarrow a}; X_{v\rightarrow a}]), H_v \\ &= Concat([X_{v\rightarrow v}; X_{a\rightarrow v}; X_{l\rightarrow v}]) \end{aligned} \tag{6}$$

In formula (6), *Concat*() means the merge operation. The extraction of features integrates the inter-sequence relationships within each modality and the inter-modality relationships between paired modalities, which effectively leverages their complementarity.

3.2. Missing modal reconstruction

After extracting multi-modal features, it is necessary to reconstruct the randomly missing modalities, thereby enabling the entire network to learn the features of these missing modalities. Previous methods employ a similar reconstruction module to improve the accuracy, but they fail to unaligned data and are less efficient in terms of running time compared to our model. In the feature reconstruction module, we leverage the relationship between missing and non-missing modalities to enhance the prediction accuracy through modality feature reconstruction. Compared with the previous methods, we reconstruct the missing data of each single modality in the multi-modal missing features reconstruction. In addition, we also calculate the loss of each modality reconstruction to strengthen the reconstruction ability. Specifically, as is shown in Fig. 3, the missing modal reconstruction module first utilizes the attention mechanism to encode the missing modalities and learn the relationships of features, and then, the reconstruction results are mapped to the dimensions of extracted features. The reconstruction of feature E^{con} is shown in formulas (7) and (8).

$$E^{con}_l = g(H_l, H_l, H_l), E^{con}_a = g(H_a, H_a, H_a), E^{con}_v = g(H_v, H_v, H_v) \tag{7}$$

$$E_l = linear(E^{con}_l), E_a = linear(E^{con}_a), E_v = linear(E^{con}_v) \tag{8}$$

In formula (7), $g(\cdot)$ means the attention mechanism, H_l, H_a and H_v means the features of language, acoustic and visual modality respectively according to extracted multi-modal features. In formula (8), *linear* means the function for linear mapping, E_l, E_a and E_v are the reconstruction features of language modality, acoustic modality and visual modality.

3.3. Modality-invariant temporal representation learning

The modality-invariant representation learning is a crucial approach for addressing the heterogeneity between different modalities. However, traditional feature-level similarity constraints modality-invariant representation loses important specific information due to the alteration of modality-specific details; Distribution-level similarity constraints modality-invariant representation directly treats each modality as an individual distribution, but it neglects the sequence information and thus reducing the effectiveness of model. Previously, there was no existing approach to address the missing modality problem using modality-invariant representation. To solve these problems, on the basis of previous research (Sun et al., 2023), we innovatively apply temporal-level similarity representation to effectively solve the fusion problem associated with missing modality and propose a novel modality-invariant representation learning approach based on temporal-level similarity constraints.

The specific implementation of modality-invariant temporal representation learning is shown as follows. To obtain the modalities specific features and reduce domain shifts, the feature information H_l, H_a and H_v extracted by missing modality input are mapped to the latent common space through linear operations to obtain Z_l, Z_a and Z_v. Our method treats each modality as a multivariate Gaussian distribution (considering each timestamp as a single Gaussian distribution) to avoid the loss of specific information of each modality. We model the time-series of each modality as a Gaussian distribution $N(\mu_w, \sigma^2_w)$ with the expectation σ and standard deviation μ. Consequently, the entire series becomes a multivariate Gaussian distribution $N(\boldsymbol{\mu}_w, \Sigma^2_w)$ with the expectation $\boldsymbol{\mu}_w$ and standard deviation Σ_w. Assuming that modalities w_1 and w_2 follow distributions $p_{w_1}(x) \sim N(\boldsymbol{\mu}_{w_1}, \Sigma^2_{w_1})$ and $p_{w_2}(x) \sim N(\boldsymbol{\mu}_{w_2}, \Sigma^2_{w_2})$ respectively, it can be expressed as formula (9).

$$\begin{aligned} p_{w_1}(x) \sim N(\boldsymbol{\mu}_{w_1}, \Sigma^2_{w_1}) &= \frac{1}{(2\pi)^{\frac{1}{2}t}|\Sigma_{w_1}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\boldsymbol{\mu}_1)^{\mathrm{T}}\Sigma^{-1}_{w_1}(x-\boldsymbol{\mu}_1)), \\ p_{w_2}(x) \sim N(\boldsymbol{\mu}_{w_2}, \Sigma^2_{w_2}) &= \frac{1}{(2\pi)^{\frac{1}{2}t}|\Sigma_{w_2}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}_{w_2}(x-\boldsymbol{\mu}_2)) \end{aligned} \tag{9}$$

And then, the KL divergence is employed to capture the implicit temporal distribution-level similarities, and modality-invariant representations are learned by computing the similarity of multivariate Gaussian distribution, which is shown in formula (10).

$$\begin{aligned} &D_{KL}(p_{w_1}(x)\|p_{w_2}(x)) \\ &= \int [log(p_{w_1}(x)) - log(p_{w_2}(x))] p_{w_1}(x) dx \\ &= \int log(p_{w_1}(x)) p_{w_1}(x) dx - \int log(p_{w_2}(x)) p_{w_1}(x) dx \end{aligned} \tag{10}$$

The two algebraic expressions in formula (10) are calculated in formula (11).

$$\int log(p_{w_1}(x))p_{w_1}(x)dx = log(\frac{1}{(2\pi)^{\frac{1}{2}t}|\Sigma_{w_1}|^{\frac{1}{2}}}) - \frac{1}{2}E_{p_{w_1}(x)}\Gamma_1,$$
$$\int log(p_{w_2}(x))p_{w_1}(x)dx = log(\frac{1}{(2\pi)^{\frac{1}{2}t}|\Sigma_{w_2}|^{\frac{1}{2}}}) - \frac{1}{2}E_{p_{w_1}(x)}\Gamma_2 \quad (11)$$

In formula (11), $\Gamma_1 = (x-\mu_{w_1})^{\mathrm{T}}\Sigma_{w_1}^{-1}(x-\mu_{w_1})$, $\Gamma_2 = (x-\mu_{w_2})^{\mathrm{T}}\Sigma_{w_2}^{-1}(x-\mu_{w_2})$.

As is shown in formula (10), the KL divergence presented above can be transformed into formula (12).

$$D_{KL}(p_{w_1}(x)||p_{w_2}(x)) = \int log(p_{w_1}(x))p_{w_1}(x)dx - \int log(p_{w_2}(x))p_{w_1}(x)dx$$
$$= \frac{1}{2}log(\frac{|\Sigma_{w_2}|}{|\Sigma_{w_1}|}) + \frac{1}{2}E_{p_{w_1}(x)}(\Gamma_2 - \Gamma_1) = \frac{1}{2}[log(\frac{|\Sigma_{w_2}|}{|\Sigma_{w_1}|}) + tr(\Sigma_{w_2}^{-1}\Sigma_{w_1}) - t] + \Gamma_3 \quad (12)$$

In formula (12), $\Gamma_3 = (\mu_{w_1} - \mu_{w_2})^{\mathrm{T}}\Sigma_{w_2}^{-1}(\mu_{w_2} - \mu_{w_1})^{\mathrm{T}}$ and $E_{p_{w_1}(x)}$ are the expectations, *tr*() is the trace function. The constrained feature is the modality-invariant representation feature. The implementation of formula (12) occurs during the training process, and the KL divergence is incorporated into the loss function. The modality-invariant representation learning module learns correlations among multi-modal data, to strengthen the consistency features across different modalities for accurate prediction categorization. In contrast with previous approaches, our method can effectively learn the consistency features for missing multi-modal data, thus improving the accuracy of the model.

3.4. Multi-modal fusion

After feature extraction, missing modalities reconstruction and modality-invariant temporal representation learning, we integrate different features of modalities and validate the model performance through classification.

The features are encoded and enhanced through the Gate encoder, followed by fusion using the vector fusion function, and the obtained features are then input into the classifier. As is shown in Fig. 3, a bidirectional GRU layer is first employed to bidirectionally calculate the input sequences with formula (13) for improving the accuracy in multi-class classification tasks. Through combining the forward and backward propagation of GRU to form a bidirectional model Bi-GRU that processes input sequences both in forwards and backwards. The outputs of forward and backward propagations are concatenated to form the output of Bi-GRU. And then, an 1D-convolution layer is designed as shown in formula (13) to further encode the input features, where the encoded results act as gates that selectively filter out irrelevant context information. Subsequently, the Hadamard product operation as shown in formula (15) is applied between the outputs of GRU layer and these gates. Finally, the linear mapping and pooling operations are performed on these output results to obtain feature outputs from Gate encoder, which is shown in formula (16).

$$F_l^{gru} = BG(H_l), F_a^{gru} = BG(H_a), F_v^{gru} = BG(H_v) \quad (13)$$

$$gate_l = Conv1d(F_l^{gru}), gate_a = Conv1d(F_a^{gru}), gate_v = Conv1d(F_v^{gru}) \quad (14)$$

$$F_l^{hp} = tanh(F_l^{gru}) \bullet sig(gate_l), F_a^{hp} = tanh(F_a^{gru}) \bullet sig(gate_a), F_v^{hp} = tanh(F_v^{gru}) \bullet sig(gate_v) \quad (15)$$

$$F_l^{lin} = linear(F_l^{gru}, F_l^{hp}), F_a^{lin} = linear(F_a^{gru}, F_a^{hp}), F_v^{lin} = linear(F_v^{gru}, F_v^{hp}) \quad (16)$$

The encoding features of different modalities are ultimately integrated as shown in formula (17).

$$F = Concat([Maxpool(F_l^{lin}); Maxpool(F_a^{lin}); Maxpool(F_v^{lin})]) \quad (17)$$

The classification outcomes of the fusion results are presented as shown in formula (18).

$$F^{cls} = Classifier(F) \quad (18)$$

3.5. Optimization

The loss function is divided into three components. The first component is the reconstruction loss. Due to the SmoothL1Loss function provides smaller penalties for outliers (such as excessively large or small values), thus, we employ the SmoothL1Loss function to improve the robustness of the model. The specific loss functions utilized in different modalities are shown in formula (19).

$$\mathscr{L}_{con}^{l} = \begin{cases} \frac{1}{2}(E_l - X^l)^2, if|E_l - X^l| < 1 \\ |E_l - X^l| - \frac{1}{2}, otherwise \end{cases},$$
$$\mathscr{L}_{con}^{a} = \begin{cases} \frac{1}{2}(E_a - X^a)^2, if|E_a - X^a| < 1 \\ |E_a - X^a| - \frac{1}{2}, otherwise \end{cases}, \quad (19)$$
$$\mathscr{L}_{con}^{v} = \begin{cases} \frac{1}{2}(E_v - X^v)^2, if|E_v - X^v| < 1 \\ |E_v - X^v| - \frac{1}{2}, otherwise \end{cases}$$

In formula (19), $\mathscr{L}_{con}^{l}$, $\mathscr{L}_{con}^{a}$, $\mathscr{L}_{con}^{v}$ represent the loss of language, acoustic and visual, respectively. And then, the construction loss is shown in formula (20).

$$\mathscr{L}_{con} = \omega_1\mathscr{L}_{con}^{v} + \omega_2\mathscr{L}_{con}^{v} + \omega_3\mathscr{L}_{con}^{v} \quad (20)$$

In formula (20), ω_1, ω_2 and ω_3 are the weight. We adjust the weight setting in the experiment. The weight parameters of language, acoustic and visual modalities are set as {5, 2, 20} for MOSI while the weight parameters are set as {1, 0.01, 0.0001} for SIMS.

Then, the similarity loss of modality-invariant modules is calculated using the DL divergence as the loss function. The calculation procedure is shown in formula (21).

$$\mathscr{L}_{sim}^{l,a} = DL(Z_l||Z_a) = \Sigma_x Z_l(x) log\frac{Z_l(x)}{Z_a(x)},$$
$$\mathscr{L}_{sim}^{v,a} = DL(Z_v||Z_a) = \Sigma_x Z_v(x) log\frac{Z_v(x)}{Z_a(x)}, \quad (21)$$
$$\mathscr{L}_{sim}^{l,v} = DL(Z_l||Z_v) = \Sigma_x Z_l(x) log\frac{Z_l(x)}{Z_v(x)}$$

The similarity loss of the three modal pairs is then calculated as shown in formula (22).

$$\mathscr{L}_{sim} = \theta_1\mathscr{L}_{sim}^{l,a} + \theta_2\mathscr{L}_{sim}^{v,a} + \theta_3\mathscr{L}_{sim}^{l,v} \quad (22)$$

In formula (22), θ_1, θ_2 and θ_3 are the weight. In the experiment, the similarity loss is taken as the average of three modality pairs and calculated as follows: $\mathscr{L}_{sim} = (\mathscr{L}_{sim}^{l,a} + \mathscr{L}_{sim}^{v,a} + \mathscr{L}_{sim}^{l,v})/3$. The weight parameters are the same.

We then use the L1Loss function as classification loss function, which is implemented as shown in formula (23).

$$\mathscr{L}_{cls} = \frac{1}{n}\sum_{i-1}^{n}|F_i^{cls} - Y_i| \quad (23)$$

Finally, the objective function is calculated using formula (24).

$$\mathscr{L} = \omega_1\mathscr{L}_{con} + \omega_2\mathscr{L}_{sim} + \omega_3\mathscr{L}_{cls} \quad (24)$$

In formula (24), ω_1, ω_2 and ω_3 are the weight. In the experiment, we set the weight parameters for reconstruction loss, similarity loss and

classification loss as {1, 0.001, 1}.

4. Experiment

In order to test the performance of proposed MIT-FRNet model, we set the following research questions (RQs) and then conduct extensive experiments on two datasets to answer these RQs.

1. **RQ1**: Does the proposed MIT-FRNet model outperform the state-of-the-art models in multi-modal prediction capability?
2. **RQ2**: Does the missing modal reconstruction (MMR) and modality-invariant temporal representation (MITR) network improve the efficiency of feature reconstruction network models in the field of random missing multi-modal?
3. **RQ3**: Does the combination of these two innovations in feature reconstruction network architecture deliver better convergence?

4.1. Datasets

To verify the efficiency of proposed MIT-FRNet model, we conduct extensive experiments on the CMU-MOSI (abbreviated as MOSI) and SIMS datasets. The basic distribution and statistics of each dataset are shown in Table 2. Both of the MOSI and SIMS dataset include three modalities: acoustic modality, language modality, and visual modality. The specific introduction of these two datasets is shown as follows:

MOSI (Zadeh, Zellers, Pincus, & Morency, 2016**):** The MOSI dataset includes the video clips of 2199 views from 93 speakers in youtube. On average, each video has 23.2 opinion segments, the average length of each opinion segment is about 4.2 s, and the total words of expressed opinions are 26,295. Each video consists of multiple opinion clips, and each clip is annotated with a sentiment in the range [−3,3], where −3 is highly negative, 3 is highly positive, and 0 is neutral. Specifically, the MOSI dataset contains 7 categories, including strongly positive (labeled as +3), positive (+2), weakly positive (+1), neutral (0), weakly negative (−1), negative (−2), strongly negative (−3).

SIMS (Yu, Xu, Meng, Zhu, Ma, Wu, Zou, & Yang, 2020**):** The SIMS dataset is composed of 60 original videos and 2281 valid video clips from movies, TV dramas and lifestyle variety shows. These videos encompass a range of facial expressions, head movements, shielding gestures and other actions with an average clip duration of 3.67 s, where 1500 are male subjects and 781 are female subjects. The included five sub-categories of emotions are as follows: intense negative (−1.0, −0.8), weakly negative (−0.6, −0.4, −0.2), neutral (0.0), weakly positive (0.2, 0.4, 0.6) and intense positive (0.8, 1.0).

4.2. Baselines

In the experiment, we mainly test the efficiency of proposed MIT-FRNet model with the use of four state-of-the-art models. The specific information of each compared model is shown as follows:

TFN (Tensor Fusion Network) (Zadeh, Chen, Poria, Cambria, & Morency, 2017**):** TFN is a novel end-to-end fusion approach for multi-modal fusion, it encompasses the comprehensive learning of intra-modal and inter-modal dynamics along with explicit aggregation of unimodal, bimodal and trimodal interactions.

MulT (Multimodal Transformer) (Tsai, Bai, Liang, Kolter, Morency, & Salakhutdinov, 2019**):** MulT effectively addresses the consistency in an end-to-end manner, where the core is directed pair-wise cross-modal attention. It specifically focuses on capturing the interaction between multi-modal sequences at different time steps and potentially adjusting the information flow across modalities.

MISA (Modality-Invariant and -Specific representations framework) (Hazarika et al., 2020**):** MISA projects each modality to two distinct subspaces. The first subspace is called modality-variant subspace, it facilitates the learning of cross-modal representations by identifying commonalities and reducing the gaps between modalities. The second subspace is modality-specific subspace, it captures unique characteristics specific to each individual modality.

TFR-NET (Transformer-based Feature Reconstruction Network) (Yuan et al., 2021**):** TFR-NET utilizes an intra-modal and inter-modal attention-based extractor to acquire a robust representation of each element in the modal sequence, it incorporates a reconstruction module to generate missing modal features, thereby improving the model resilience towards random omissions in unaligned modal sequences.

Table 2
Dataset distribution and statistics for benchmark dataset in format negative/neutral/positive.

Dataset	# Train	# Valid	# Test	# All
MOSI	552/53/679	92/13/124	379/30/277	2199
SIMS	742/207/419	248/69/139	248/69/140	2281

4.3. Evaluation metrics

In the experiment, we evaluate the performance of proposed MIT-FRNet model for all missing rates in terms of binary classification accuracy (Acc-2), five classification accuracy (Acc-5), F1-Score (F1), Mean Absolute Error (MAE), and Pearson Correlation coefficient I. The detailed information of these five widely-used metrics are presented in Table 3, where TP_i is the number of predicting the i-th category as the i-th category, FP_i is the number of predicting the j-th category as the i-th category ($i \neq j$), TN_i is the number of predicting the j-th category as the j-th category ($i \neq j$), FN_i is the number of predicting the i-th category as the j-th category ($i \neq j$). The higher indicators of Acc-I (I classification accuracy), F1 and C means the better performance of the model. The lower MAE indicator means the better performance of the model. Additionally, the MOSI dataset is used to further verify the accuracy of binary classification (Acc-2) and seven classification (Acc-7), the SIMS dataset is used to further verify the accuracy of three classification (Acc-3), thereby enriching the analysis of proposed method.

4.4. Modality preprocessing

The above two datasets used in the experiments conclude three modalities, including language modalities, acoustic modalities and visual modalities. For each of the three modalities, we process the information from videos as follows. Furthermore, we have set the missing rate to indicate the proportion of missing modal data and the mask to mark the data as invalid during the data processing. The missing rate

Table 3
Evaluation metric.

Metric	Formula	Meaning
Accuracy	$A = \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i}$	The proportion of correctly predicted samples. ($i = 2, 3, 5, 7$)
F1 Score	$F1 = \frac{2^* \frac{TP_i}{TP_i + FP_i} \frac{TP_i}{TP + FN_i}}{\frac{TP_i}{TP_i + FP_i} + \frac{TP_i}{TP + FN_i}}$	A metric of considering both of precision (The proportion of correctly true samples to correctly predicted samples) and recall (The proportion of correctly predicted samples to true correctly samples)
Mean Absolute Error	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$	The mean absolute error of true samples (y_i) and predicted samples ($\hat{y}_i$)
Pearson Correlation coefficient	$C = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$	The correlation between two variables (x_i, y_i) in a classification

value is the range of {0.0, 0.1,0.2, 0.3,0.4,0.5,0.6,0.7, 0.8,0.9,1.0}. Data in the dataset are missing randomly according to the missing rate. For instance, for the dataset MOSI, with a missing rate of 0.3, the mask marks 30 percent of the modal data as invalid with a special marker. This is used to simulate the missing data in real data.

Language representation: The pre-trained Bert (Devlin, Chang, Lee, & Toutanova, 2018) is used to extract text features for each text utterance, thereby obtaining a 768-dimensional word text vector.

Acoustic representation: The audio analysis toolkit COVAREP acoustic framework (Degottex, Kane, Drugman, Raitio, & Scherer, 2014) and Librosa (McFee, Raffel, Liang, Ellis, McVicar, Battenberg, & Nieto, 2015) are used to extract the acoustic features for MOSI and SIMS separately, where three features of zero-crossing rate, mel frequency cepstral coefficient (MFCC) and constant Q Togram (CQT) are selected to represent the audio segment at 16000 Hz, and then they are connected to generate 5-dimensional acoustic features for MOSI and 33-dimensional acoustic features for SIMS.

Visual representation: The Facet and OpenFace2.0 toolkit (Baltrusaitis, Zadeh, Lim, & Morency, 2018) are used to extract the facial features for MOSI and SIMS datasets, thereby obtaining 5-dimensional and 709-dimensional visual representation respectively, including face, head and eye movements.

4.5. Experimental results

4.5.1. Answer to RQ1

To answer RQ1, we conduct extensive experiments on datasets MOSI and SMIS to verify the detection efficiency of proposed MIT-FRNet with the comparison of four state-of-the-art models, including TFN (Zadeh et al., 2017), MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020) and TFR-Net (Yuan et al., 2021). The parameter settings for these compared methods are consistent with those reported in the original papers. The previous studies on missing modalities have not considered a comprehensive assessment of the missing rate, in our experiments, we considered all possible missing rates (missing rate = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0) to verify whether the proposed MIT-FRNet can obtain better performance for various scenarios. The experimental results are shown in Tables 4 and 5.

Table 4
Experimental results of different missing rates on MOSI dataset (%).

Metrics	baselines	Missing rate										
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Acc-2 (·)	TFN	77.55	74.49	69.68	65.60	62.88	55.44	55.15	52.14	51.65	51.41	47.45
	MulT	80.52	76.63	72.59	70.50	66.96	62.24	59.33	56.17	53.40	53.76	45.92
	MISA	81.63	78.72	76.05	72.93	71.47	64.25	64.17	56.61	44.75	44.75	44.75
	TFR-Net	81.15	79.59	76.92	71.43	71.62	66.81	64.19	59.82	60.20	52.72	51.10
	Ours	**82.07**	**80.09**	**77.26**	**72.97**	**71.67**	**68.71**	**64.29**	**63.75**	**60.72**	**54.28**	**51.56**
F1 (·)	TFN	77.58	74.43	69.64	65.82	63.49	58.80	55.79	56.43	54.73	54.10	58.15
	MulT	80.56	76.71	72.54	70.81	68.07	62.48	59.50	56.52	54.47	55.77	60.05
	MISA	81.75	78.71	76.05	72.98	71.53	65.73	63.88	61.11	61.03	57.83	60.03
	TFR-Net	81.19	79.56	77.00	73.24	71.78	66.94	64.46	61.03	61.45	57.00	58.36
	Ours	**82.07**	**80.10**	**77.23**	**73.98**	**71.96**	**68.90**	**64.66**	**63.81**	**61.85**	**57.90**	**60.10**
Acc-2 (*)	TFN	78.61	75.36	70.27	65.85	63.31	54.27	54.62	50.46	50.31	50.15	48.27
	MulT	82.01	77.79	73.73	72.05	68.24	63.06	59.25	55.44	52.29	53.01	43.45
	MISA	83.50	80.29	77.13	74.29	72.31	64.23	65.70	55.69	42.23	42.63	44.62
	TFR-Net	82.52	80.90	78.20	73.93	72.66	66.92	64.63	59.30	59.55	51.02	46.36
	Ours	**83.53**	**81.45**	**78.35**	**74.75**	**73.31**	**69.87**	**67.97**	**64.08**	**59.65**	**53.15**	**47.09**
F1 (*)	TFN	78.58	75.22	70.14	65.96	63.80	57.46	55.10	54.54	53.20	52.66	56.15
	MulT	81.99	77.81	73.59	72.23	69.17	63.18	59.29	55.59	53.13	54.60	58.48
	MISA	83.41	80.56	77.07	74.24	72.16	65.58	65.84	59.37	59.38	**59.38**	59.38
	TFR-Net	82.5	80.80	78.18	74.07	72.19	66.92	64.79	60.37	60.56	55.07	59.11
	Ours	**83.48**	**81.39**	**78.25**	**74.63**	**72.35**	**69.95**	**67.84**	**64.01**	**60.65**	55.67	**60.27**
Acc-5	TFN	39.17	34.94	30.42	26.87	25.32	21.96	17.93	17.54	18.08	17.78	16.18
	MulT	42.86	36.25	32.12	28.23	25.61	24.40	20.07	19.15	17.83	16.67	15.65
	MISA	47.42	42.47	40.01	37.46	27.16	25.12	23.13	18.76	15.60	15.45	15.45
	TFR-Net	50.88	40.28	39.31	36.06	31.49	26.24	25.61	24.10	21.52	17.74	15.60
	Ours	**52.87**	**46.35**	**40.23**	**37.56**	**33.63**	**29.59**	**26.74**	**25.22**	**21.53**	**17.96**	**16.13**
Acc-7	TFN	33.92	31.20	28.23	25.17	24.10	20.89	17.54	17.20	17.69	17.40	16.18
	MulT	37.32	32.27	28.91	27.11	24.78	23.71	19.78	19.15	17.74	16.67	15.65
	MISA	41.69	37.90	35.69	33.82	25.95	24.15	22.45	18.66	15.60	15.45	15.45
	TFR-Net	42.91	35.76	34.89	33.77	29.20	24.88	23.52	23.37	21.04	17.64	15.60
	Ours	**45.29**	**38.63**	**35.81**	**34.37**	**30.32**	**27.65**	**24.53**	**24.34**	**21.80**	**17.96**	**16.13**
MAE	TFN	96.50	106.82	116.80	127.26	130.38	144.89	141.74	148.18	147.19	148.45	148.55
	MulT	88.31	97.60	108.93	116.03	124.82	133.45	136.41	140.79	142.93	142.12	147.81
	MISA	76.74	85.42	94.69	99.42	113.71	121.08	127.81	143.96	148.22	147.59	148.01
	TFR-Net	76.93	88.10	93.62	99.38	110.15	124.37	126.62	133.22	138.54	145.42	145.73
	Ours	**74.52**	**85.03**	**92.18**	**99.27**	**108.32**	**115.61**	**123.58**	**130.99**	**138.48**	**142.87**	**145.65**
C	TFN	65.67	59.13	49.40	41.47	36.89	26.20	15.27	17.13	12.35	10.58	7.63
	MulT	72.80	64.37	56.29	49.36	38.75	27.39	20.54	16.30	13.43	14.07	8.00
	MISA	77.37	73.33	66.49	60.35	52.44	46.61	38.55	24.08	9.76	6.68	–
	TFR-Net	78.18	73.72	67.61	62.67	52.88	45.12	37.21	34.10	26.40	13.55	1.46
	Ours	**78.60**	**73.75**	**67.86**	**62.74**	**53.70**	**47.38**	**38.80**	**34.12**	**28.48**	**14.39**	–

Table 5
Experimental results of different missing rates on SMIS dataset (%).

metrics	baselines	Missing rate										
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Acc-2	TFN	65.55	66.22	68.71	69.62	69.34	69.42	**71.55**	69.27	69.19	68.64	69.37
	MulT	67.37	66.00	69.37	68.64	69.37	69.37	69.37	69.37	69.37	69.37	69.37
	MISA	67.67	66.37	69.51	69.37	69.37	69.29	69.44	69.29	69.37	69.37	69.37
	TFR-Net	68.10	65.13	68.34	69.44	69.37	68.17	66.96	69.27	69.37	68.71	69.37
	Ours	**68.20**	**67.10**	**69.58**	**69.64**	**69.72**	**69.45**	69.22	**69.37**	**69.37**	**69.37**	**69.39**
Acc-3	TFN	40.84	45.14	50.47	52.72	48.99	54.75	52.29	**55.09**	52.16	47.12	54.19
	MulT	43.03	39.53	54.27	49.96	48.77	54.12	54.27	54.19	42.09	53.24	54.27
	MISA	44.27	43.76	52.81	53.54	47.38	53.17	52.15	52.74	52.12	53.17	54.27
	TFR-Net	46.90	42.01	49.89	46.68	47.19	40.04	41.50	53.46	52.12	48.58	50.47
	Ours	**47.85**	**49.45**	**54.78**	**52.74**	**54.10**	**55.17**	**55.59**	53.41	**52.37**	**55.14**	**54.27**
Acc-5	TFN	22.03	23.46	23.04	21.42	25.46	21.66	21.59	23.38	23.08	22.25	**21.30**
	MulT	21.95	19.11	21.30	17.01	21.23	21.30	21.23	22.39	19.33	21.30	21.23
	MISA	21.44	21.23	20.86	17.30	21.15	20.85	20.79	21.08	21.15	21.23	21.23
	TFR-Net	21.44	21.37	22.60	20.57	22.61	19.69	20.04	21.90	22.32	**22.47**	20.23
	Ours	**23.85**	**23.92**	**23.05**	**21.44**	**26.26**	**21.95**	**21.60**	**23.60**	**23.19**	22.01	20.64
F1	TFN	70.96	70.23	68.36	72.12	71.83	73.76	75.52	74.17	75.20	70.9	81.91
	MulT	80.14	73.75	79.91	80.45	81.91	81.91	81.41	81.91	81.91	81.91	81.91
	MISA	**81.91**	74.40	79.87	81.88	81.78	81.67	81.46	81.80	81.91	81.91	81.91
	TFR-Net	75.47	69.35	74.22	79.08	79.09	81.91	74.45	81.91	81.91	80.13	81.90
	Ours	76.97	**74.47**	**79.92**	**81.89**	**81.91**	**81.92**	**81.56**	**81.91**	**81.91**	**81.91**	**81.91**
MAE	TFN	56.87	57.41	56.60	57.58	56.54	**52.71**	**53.38**	57.60	58.40	58.32	59.84
	MulT	58.13	58.55	58.53	58.57	59.07	59.04	59.32	58.55	59.36	58.77	59.12
	MISA	58.99	59.12	59.17	59.34	59.41	59.44	59.30	59.40	59.27	59.43	59.45
	TFR-Net	56.42	57.25	56.60	57.52	57.75	58.80	59.54	58.05	58.69	58.32	59.04
	Ours	**56.07**	**56.54**	**56.39**	**57.48**	**56.18**	57.95	58.40	**57.16**	**57.90**	**58.18**	**59.03**
C	TFN	24.84	19.26	27.30	12.30	**35.33**	**38.98**	**38.82**	**39.81**	**34.12**	9.02	1.07
	MulT	10.25	11.46	10.45	6.65	3.66	6.34	3.28	9.23	7.21	8.45	0.54
	MISA	10.31	10.01	8.60	3.92	6.00	5.80	6.79	5.67	6.67	–	–
	TFR-Net	24.53	15.59	21.26	12.28	10.99	7.56	7.04	11.85	4.99	7.63	2.78
	Ours	**25.05**	**19.28**	**27.38**	**12.31**	24.75	15.94	9.56	15.73	7.46	**9.42**	–

It can be seen from Table 4 that on the MOSI dataset, the proposed MIT-FRNet has the best performance under various missing rates, it is owing to that MIT-FRNet enhances the feature by encoding missing modalities through attention mechanism as well as achieves a more refined modality-invariant temporal representation learning, leading to further enhancement of model classification performance compared to benchmark model.

As is shown in Table 5, the Pearson Correlation coefficient (C) of MIT-FRNet is suboptimal due to features loss caused by higher missing rates. When the missing rate is set to 0.9 and 1.0, the MISA fails to obtain the Pearson Correlation coefficient; When the missing rate is set to 1.0, MIT-FRNet fail to obtain the Pearson Correlation coefficient; It is owing to that these models are unable to perform correlation analysis due to the sparsity of features. For the TFN method, it achieved the best results on ACC-2 when the missing rate is set to 0.6, it achieved the best results on ACC-3 when the missing rate is set to 0.7, it achieved the best results on ACC-5 when the missing rate is set to 1.0. The MIT_FRNet model is comparatively inferior to those achieved by the TFN benchmark in a few missing rates, this discrepancy can be primarily attributed to the fact that the SIMS dataset possesses highly detailed labels with corresponding modalities, whereas the fusion principle employed by the TFN model aims at integrating all features together, this fusion strategy enhances the granularity of each category features and consequently improves classification accuracy. However, it also leads to exponential increases in computational efficiency as more labels are introduced. In addition, TFR-Net achieves the best results on Acc-5 when the missing rate is set to 0.9, MISA achieves the best results on F1 score when the missing rate is set to 0.0, they are attributed to the experimental contingency. Nevertheless, experimental results demonstrate that MIT-FRNet overwhelmingly outperforms other approaches on all six metrics. The experiment on the SMIS dataset also demonstrates that MIT-FRNet outperforms state-of-the-art methods.

Therefore, from the results, it can be seen that MIT-FRNet has comprehensively excellent predictive ability at all missing rates. This is because the model focuses on the common features between missing and non-missing modalities when reconstructing the missing modalities, which enhances the information of the missing modalities. At the same time, the modality-specific information of each modality is weakened and the modality-invariant information is enhanced. And the modality-invariant information provides more accurate feature information for the model prediction.

Besides the evaluation metrics including accuracy, F1 score, MAE and C, we also use the missing weighted average (*MWA*) to measure the performance of compared models. The calculation of *MWA* is shows in formula (25).

$$MWA(item) = \frac{1}{I}\sum_{i=0}^{I}(V_i \times (1 - i \times 10\%)), I = [0, 10] \tag{25}$$

In formula (25), *item* is a set of evaluation metrics, including accuracy/F1/MAE/C, V_i is the value of *item* evaluation metrics in the missing rates, i*10 % represents varying rates of missing data.

The experimental result on metric *MWA* is shown in Fig. 5, and the experimental result across various datasets is shown in Fig. 6 and Fig. 7.

For the proposed MIT-FRNet model, it can obtain optimal performance on *MWA* metric on both MOSI and SIMS datasets for the binary classification results as shown in Fig. 5(a), it can achieve satisfactory

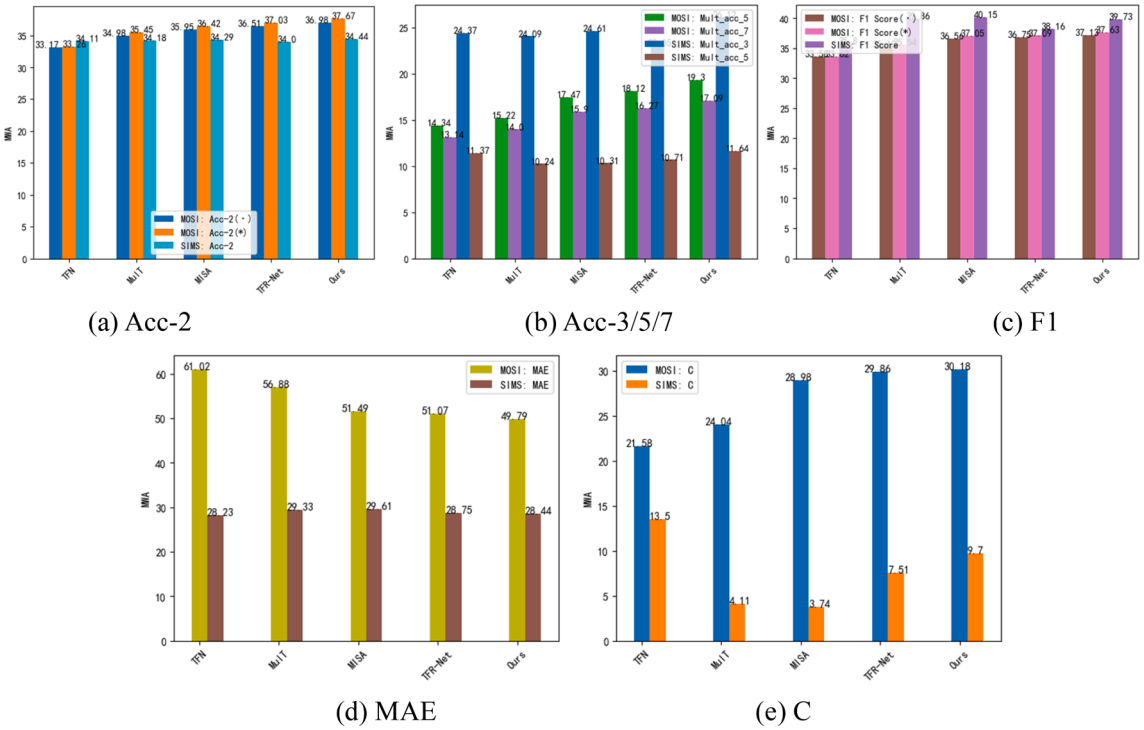

Fig. 5. The results of *MWA*.

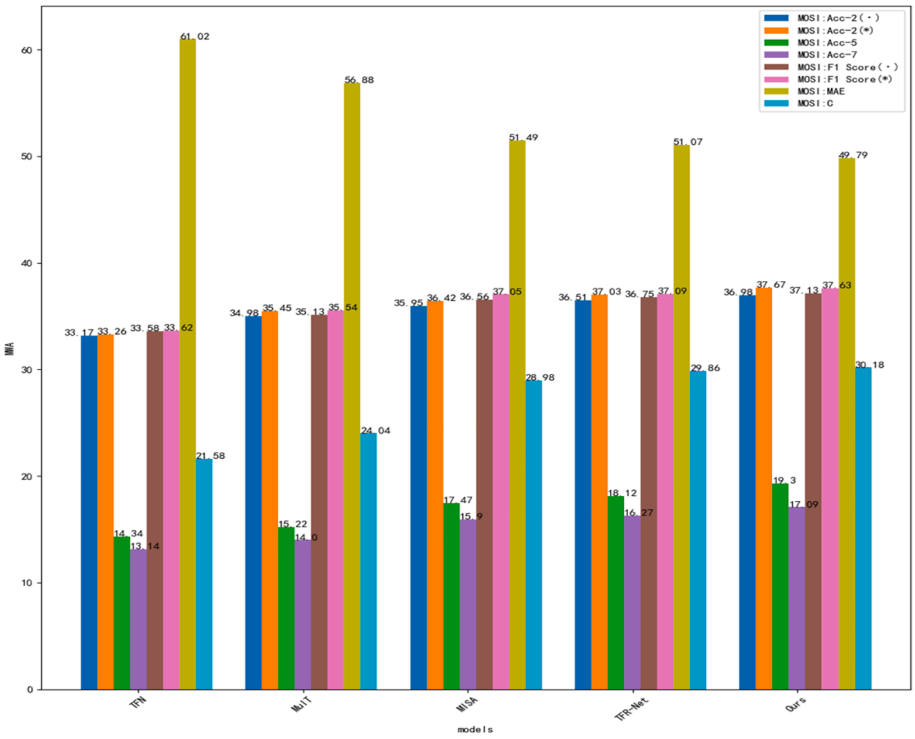

Fig. 6. All metrics *MWA* on MOSI dataset.

results on two datasets for multi-classification as illustrated in Fig. 5(b). As is shown in Fig. 5(c), in terms of F1 Score, the proposed MIT-FRNet achieves superior performance on the MOSI dataset, but the performance on SIMS dataset is inferior to that on MISA. As is shown in Fig. 5 (d), with respect to MAE, the proposed MIT-FRNet performs well on dataset MOSI, and the performance on SIMS is higher than that on MOSI. Compared with correlation coefficient (shown as Fig. 5(e)) obtained by other models, the proposed MIT-FRNet demonstrates a leading advantage when applied to MOSI dataset, but it lacks advantage slightly to TFN on SIMS dataset.

Overall, the proposed MIT-FRNet has demonstrated significant advantages on the MOSI and SIMS datasets. We observe sub-optimal results in certain evaluation metrics of *MWA* value on the SIMS dataset, this discrepancy can be primarily attributed to the fact that the SIMS dataset

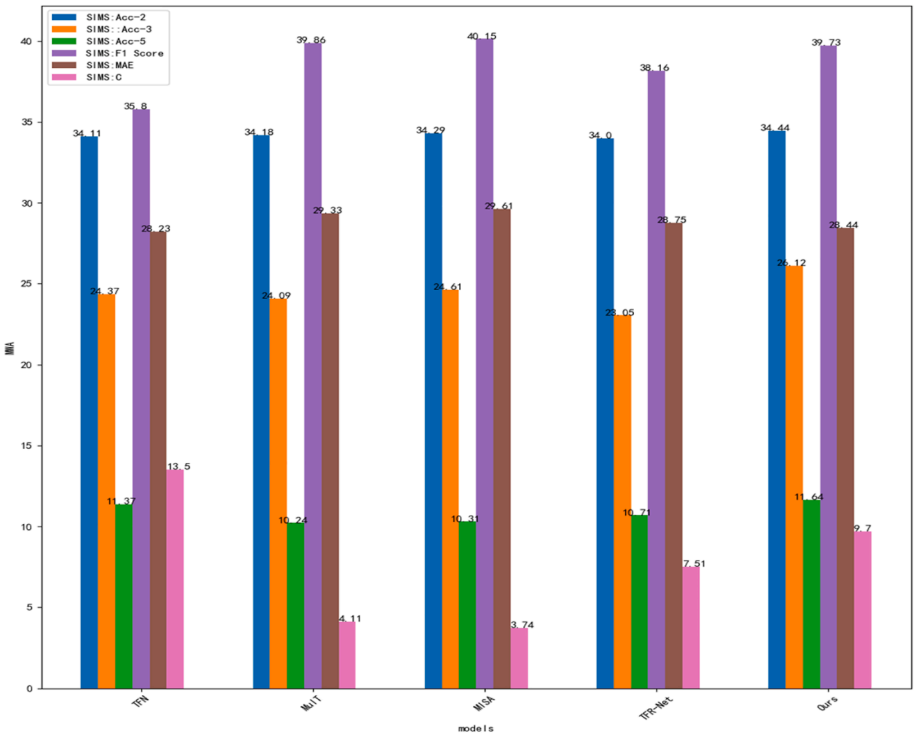

Fig. 7. All metrics *MWA* on SIMS dataset.

possesses highly detailed labels with corresponding modalities. After careful consideration of all factors, the MIT-FRNet is deemed to be the optimal choice. The *MWA* evaluation metric demonstrates the average prediction performance of proposed model under missing modalities. The experimental results show that the MIT-FRNet also has the best average prediction performance.

In addition, the running time of models is also focused and the experiment results are shown in Table 6. The TFN and MISA models focus on aligned multi-modal data, whereas the MulT, MIT-FRNet and TFR-Net models have the capability to predict both aligned and unaligned multi-modal data.

The results presented in Table 6 demonstrate that the computational time required for the models (TFN and MISA) capable of solving aligned sequences is comparatively lower than that for the models (MIT-FRNet, MulT, and TFR-Net) designed to handle unaligned sequences. The running time of the models limited to aligned multi-modal data is generally shorter compared to the models capable of handling unaligned multi-modal data, which is owing to that the attention mechanism handling unaligned data in the models leads to an increase in parameters within the network model, consequently necessitating computational resources and time. Among the models addressing unaligned sequences, our proposed MIT-FRNet model exhibits minimum running time while achieving superior accuracy. The TFN is an end-to-end fusion method that can effectively leverage tensor features extracted from the unimodal, bimodal and trimodal interaction. The MISA method is designed to enhance modal consistency and reduce multi-modal heterogeneity by projecting different modalities into subspaces, it allows for the learning of consistency characteristics while weakening the specific information of each modality. However, these two models only applicable to the aligned multi-modal data, and are unable to handle unaligned multi-modal data. The MulT is an end-to-end fusion approach, its cross-modal feature extraction module is designed for unaligned data. MulT can capture the features from multi-modal regardless of aligned or unaligned data. The TFR-Net learns semantic-level features corresponding to missing features from missing modalities for unaligned data by extraction and reconstruction modules, thereby improving the accuracy of prediction. However, the excessive amount of time they spend is noteworthy. The proposed MIT-FRNet model utilizes extractors based on intra-modal and inter-modal attention mechanisms to acquire a robust representation of each element in the multi-modal sequence, enabling feature reconstruction and obtaining a more refined temporal-level modal invariant representation. It is applicable for unaligned and aligned multi-modal data, and also provides the advantage of less running time for unaligned data.

In summary, the proposed MIT-FRNet outperforms the MulT and TFR-Net models. Although it slightly lags behind the TFN and MISA models in running time efficiency, our MIT-FRNet model has the distinct advantages: (1) It is effectively suitable for the aligned and unaligned multi-modal data; (2) It achieves superior prediction accuracy; (3) It successfully tackles random missing modalities. Conversely, both TFN and MISA models fall short in these three aspects, resulting in significantly lower performance compared to our model. Thus, our model

Table 6
The running time experiments on the MOSI dataset.

Metrics	Type	baselines	Missing rate										
			0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Running time (s)	aligned	TFN	3	0	3	1	0	3	1	1	2	1	1
		MISA	3	2	2	1	1	3	2	1	3	2	2
	unaligned	MulT	27	28	29	28	27	28	29	27	29	28	27
		TFR-Net	16	15	16	17	15	16	17	16	15	17	16
		Ours	**14**	**15**	**13**	**13**	**15**	**14**	**13**	**14**	**13**	**14**	**13**

attains optimal performance.

In our model, we use the following manners to improve the accuracy: (1) We propose an intra-modal and inter-modal attention extractor for unaligned data, while simultaneously leveraging feature information within and across modalities to capture time-level sequence features effectively; (2) We introduce a feature reconstruction network that captures missing and non-missing modal relationships, aiming to minimize modal losses as much as possible; (3) We consider modal consistency by learning the characteristics that ensure consistency between multiple modalities, thereby reducing heterogeneity among different modalities. Meanwhile, we also adopt the following manners to reduce the time consumption: (1) The attention mechanism in feature extraction and reconstruction modules focuses the model on the key features and reduces the dependence of irrelevant features, thereby simplifying the complexity of the model and decreasing time consumption; (2) Employing appropriate optimizer and parameter adjustment further improves the running speed. Extensive experimental results show that our model improve the prediction accuracy and spend less time for unaligned multi-modal data compared with baselines.

Answer to RQ1:

Extensive experimental results show that the proposed MIT-TRNet model has better multi-modal prediction capability compared to the state-of-the-art models. Feature reconstruction is based on the intra-modal and inter-modal feature extraction, taking full account of the common feature information between missing and non-missing modalities. And modality of each timestamp is considered as a Gaussian distribution and the implicit temporal similarity features are captured, thereby learning modality-invariant representations. Therefore, MIT-FRNet has strong predictive ability and fast prediction speed for aligned and unaligned multi-modal data.

4.5.2. Answer to RQ2

To answer RQ2, we conduct a series of ablation experiments on MOSI dataset to verify whether the missing modal reconstruction (MMR) and modality-invariant temporal representation (MITR) network improve the efficiency of feature reconstruction network models in the field of random missing multi-modal, and the experimental result is shown in Table 7.

For the proposed MIT-FRNet model without MMR, when the missing rate is set to 0.0 (for Acc-2 (·), F1 Score (·), Acc-2 (*)), 0.1 (all metrics), 0.3 (other metrics except for Acc-5), the experimental results show a better performance, the reason is that the MMR has less impact on the model compared with MITR when the missing rate is set lower. However, when the missing rate is set larger, the ablation results are worse than that of proposed MIT-FRNet model. Additionally, the ablation model is unable to analyze correlation due to the sparsity of features. The reconstruct loss of MMR can adjust the model parameters and gradually reduce the value of the loss function so that the model can predict the target variable more accurately. It effectively captures the correlation between missing and non-missing modalities. Thus, MMR module strengthens the predictive ability of the model.

For the model without MITR, when the missing rate is set to 0.1 (Acc-2(·), Acc-7, MAE), 0.2 (Acc-5), 0.4 (Acc-2(·), F1 Score (·)), 0.5 (F1 Score (*), Acc-5), 0.6 (other metrics except for Acc-5), 0.7 (F1 Score (*)), 0.8 (Acc-2(*), Acc-5, Acc-7), 0.9 (F1 Score (*), MAE) and 1.0 (F1 Score (·), F1 Score (*)), the ablation results illustrate that the MITR has negative influence for the MIT-FRNet model, which is caused by that the modality-invariant representations are failed or obtained less in some scenarios. MITR module captures the implicit temporal distribution-level similarities, thereby learning modality-invariant representations. It learns finer-grained feature information. Thus, The MITR module has a facilitating effect on the predictive ability of the model.

In summary, the removal of sub-network models for missing modal reconstruction and modality-invariant temporal representation has detrimental impacts on feature acquisition.

Answer to RQ2:

The experimental results show that the missing modal reconstruction and modality-invariant temporal representation network improve on feature reconstruction network models in the field of random missing multi-modal. MMR makes full use of the common information of modalities. The absence of the MMR module has a negative effect on the model. MITR learning can enhance the special information of different modalities and alleviate the heterogeneity between different modalities, thus improving the predictive ability of the model. The absence of MITR module weakens the predictive ability of the model.

4.5.3. Answer to RQ3

To answer RQ3, we conduct extensive experiments to test whether or not the combination of these two innovations to the feature reconstruction network architecture can deliver convergence, and the experimental result is shown in Fig. 8 and Fig. 9. In the experiments, we assess the model convergence on two datasets with the missing rate ranging from 0.0 to 1.0 at intervals of 0.1.

Table 7
The ablation experiments on the MOSI dataset.

	Missing rate	Acc-2 (·)↑	F1 Score (·)↑	Acc-2 (*)↑	F1 Score (*)↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑
MIT-FRNet (ablation MMR)	0.0	82.12△↑	82.27△↑	84.09△↑	84.17△↓	50.09△↓	43.34△↓	76.17△↑	77.88△↓
	0.1	80.32△↑	80.33△↑	81.50△↑	81.45△↑	48.06△↑	42.08△↑	79.74△↓	74.95△↑
	0.2	77.02△↓	77.02△↓	78.40△↑	78.33△↑	42.32△↑	36.78△↑	91.24△↓	67.83△↓
	0.3	75.22△↑	75.22△↑	75.46△↑	75.36△↑	37.56△↓	32.94△↓	100.22△↓	62.82△↑
	0.4	70.17△↓	70.86△↓	70.94△↓	71.47△↓	30.52△↓	28.43△↓	112.19△↑	54.18△↑
	0.5	66.96△↓	67.05△↓	67.02△↓	66.97△↓	24.88△↓	23.57△↓	123.28△↑	44.47△↓
	0.6	63.26△↓	64.41△↓	62.80△↓	63.80△↓	24.01△↓	22.79△↓	133.38△↑	39.87△↑
	0.7	63.51△↓	63.53△↓	63.87△↓	63.77△↓	25.22△↓	24.05△↓	126.37△↓	36.05△↓
	0.8	54.42△↓	58.96△↓	52.84△↓	57.72△↓	19.05△↓	18.51△↓	143.66△↑	23.79△↓
	0.9	54.86△↑	57.15△↓	53.71△↑	55.80△↑	15.69△↓	15.69△↓	140.94△↓	15.89△↑
	1.0	44.75△↓	61.83△↑	42.23△↓	59.38△↓	15.69△↓	15.69△↓	147.17△↑	–
MIT-FRNet (ablation MITR)	0.0	81.20△↓	81.21△↓	82.62△↓	82.58△↓	49.90△↓	43.15△↓	76.59△↑	77.80△↓
	0.1	79.35△↑	79.00△↓	79.83△↓	79.71△↓	46.02△↓	40.04△↑	84.29△↓	72.78△↓
	0.2	76.77△↓	76.72△↓	77.64△↓	77.52△↓	41.69△↑	35.76△↓	92.04△↑	67.47△↓
	0.3	73.37△↓	73.55△↓	73.38△↓	73.44△↓	34.84△↓	30.08△↓	105.48△↑	61.03△↓
	0.4	71.91△↑	71.90△↑	71.90△↓	71.78△↓	30.47△↓	28.09△↓	111.41△↑	53.40△↓
	0.5	69.15△↓	69.34△↑	69.87△↓	69.95△↓	30.81△↑	27.18△↓	116.95△↑	47.29△↓
	0.6	64.77△↑	64.02△↓	64.84△↑	64.96△↑	27.06△↑	24.93△↑	125.14△↓	40.16△↑
	0.7	61.52△↓	62.10△↑	61.08△↓	61.50△↓	24.00△↓	22.60△↓	132.90△↑	33.74△↓
	0.8	60.71△↓	60.57△↓	61.79△↑	62.31△↓	23.95△↑	23.08△↑	134.40△↑	28.41△↓
	0.9	52.04△↓	57.73△↓	50.71△↓	56.27△↑	16.37△↓	16.12△↓	148.38△↓	16.66△↑
	1.0	44.75△↓	61.83△↑	42.23△↓	59.38△↑	16.08△↓	16.08△↓	150.41△↑	2.32△↓

Explanation: △↓ means decrease after ablation. △↑ means improvement after ablation. ↑ means the higher, the better. ↓ means the lower, the better.

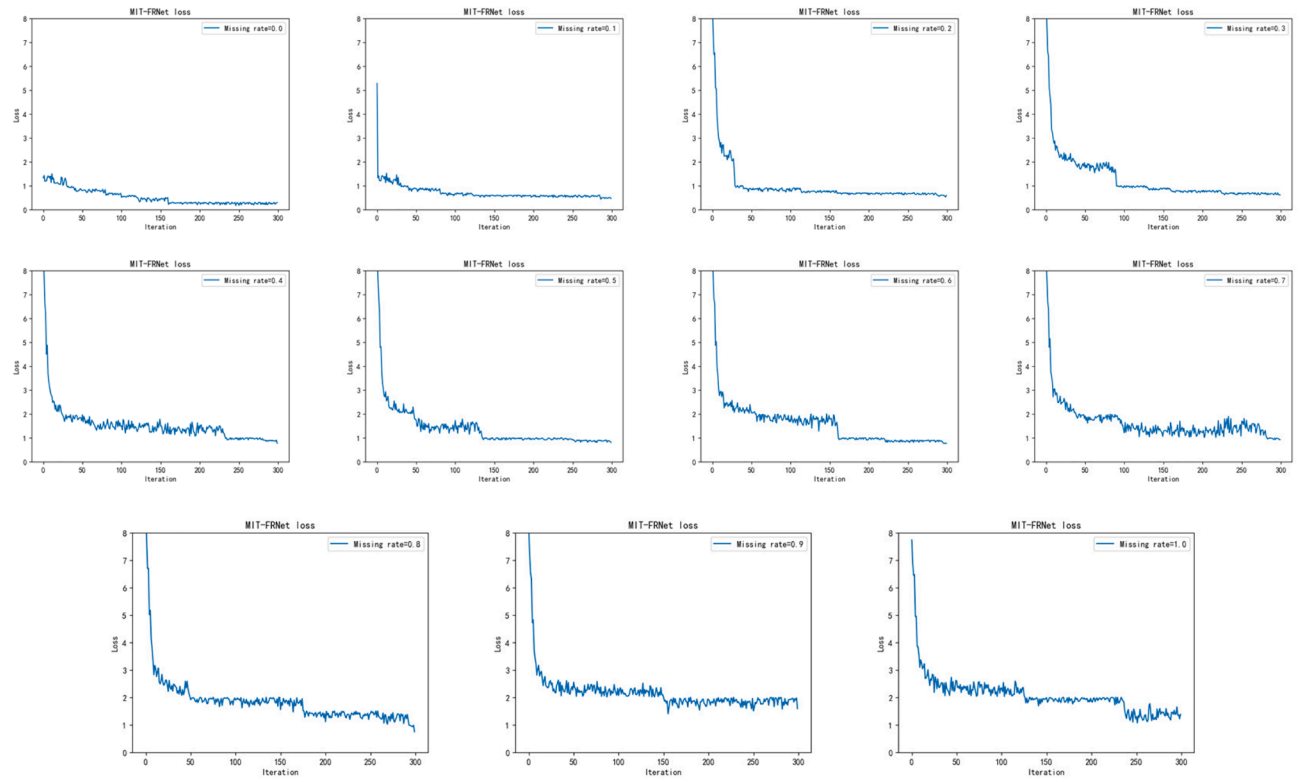

Fig. 8. Convergence visualization on MOSI dataset.

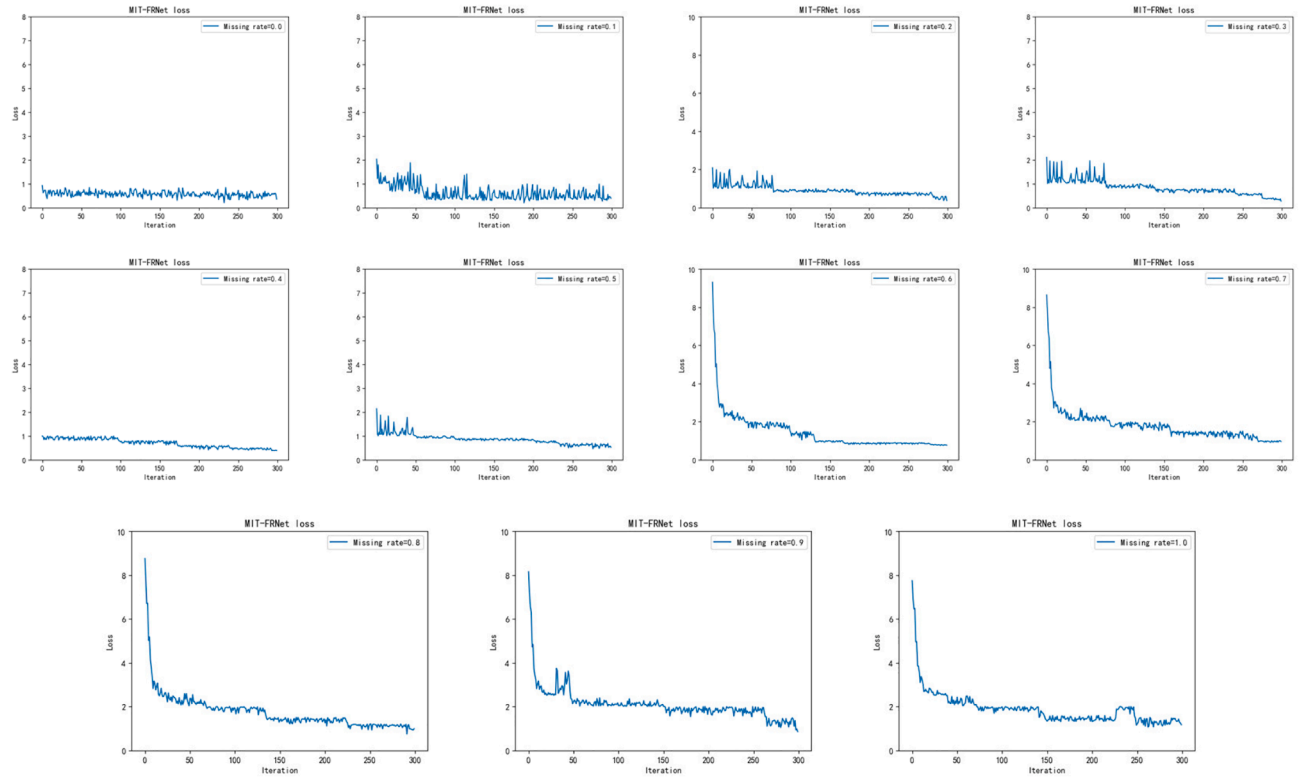

Fig. 9. Convergence visualization on SIMS dataset.

As is shown in Fig. 8, on the MOSI dataset, when the missing rate is set to 0.0, the curve is stable after iteration exceed 150; When the missing rate is set to 0.1–0.6, the loss decreases dramatically as the increasing of iterations and then change slightly; When the missing rate is set to 0.7–1.0, the loss decreases dramatically and then drops again; The experimental result shows that the proposed MIT-FRNet model has

good convergence. It can be known in Fig. 9 that on the SIMS dataset, when the missing rate is set to 0.0 and 0.4, the change of loss is minimal but the loss exhibits significant fluctuations; When the missing rate is set to 0.2, 0.5 and 0.6, the curve drops first and then becomes stable; When the missing rate is set to 0.7–1.0, the curve also drops first and then gradually stable. It is evident that the proposed MIT-FRNet model exhibits robust convergence under different missing rates during training, demonstrating its better convergence. Important modules in the MIT-FRNet, missing modal reconstruction (MMR) and modality-invariant temporal representation (MITR) network, are set up with reasonable loss functions. The former loss function captures the information between missing and non-missing modal common features, and the latter constrains the consistent features and information between different modalities. This indicates that the MIT-FRNet learns the key features of the modalities data faster during training. By comprehensive model architecture and learning the crucial features in different modalities data, the model is convergent and able to implement better predictions.

Answer to RQ3:

The experimental results show that the combination of these two innovations to the feature reconstruction network architecture delivers convergence. The reasonable model architecture and loss functions help the model to better converge to a suitable solution by adjusting the parameters during the training.

5. Conclusion

The research of missing modalities is one of the key fields in the multi-modal learning, however, previous studies only focused on the fusion scheme to improve classification performance but failed to improve the robustness in missing modalities representation as well as consider the heterogeneity in different modalities. In this paper, we propose a modality-invariant temporal representation learning-based feature reconstruction network called MIT-FRNet for missing modalities, which aims to enhance the robustness of multi-modal representation and effectively learn modality-invariant representation, thereby mitigating the heterogeneity between modalities and further addressing the challenges of incomplete multi-modal sequences. Specifically, the multi-modal feature extraction module is used firstly to extract the latent features from different modalities. Secondly, the missing modal reconstruction module employs Transformer encoder to reconstruct missing modalities features. Thirdly, the modality-invariant temporal representation learning module introduces more fine-grained similarity constraints to learn modality-invariant temporal representations. Finally, multi-modal fusion module fuses and classifies predicted features from different modalities. Compared with the state-of-the-art methods, extensive experiments demonstrate that the proposed MIT-FRNet performs better classification performance (with higher accuracies, higher F1 scores, lower MAE, higher correlation) and more stable convergence.

The MIT-FRNet model is applied in the field of emotion recognition, and its applicability extends to the medical domain for the treatment of psychological disorders. In medical applications, doctors can assess the type and severity of a patient's illness based on the acoustic, language, and visual information provided by the patient. Similarly, the model can be employed in anomaly detection or security domains. Analyzing facial expressions, language, and other cues enables the system to identify abnormal situations, thereby enhancing security. In addition, it can also be extended to the field of education, media and so on.

Although the MIT-FRNet can achieve better performance, but it does not consider the data imbalance problem. In future work, we would like to introduce diffusion models or generative adversarial networks (GANs) to enhance minority class data, thereby solving the data imbalance problem. In addition, we also would like set different weights for diverse modalities and relies on multi-modal translation, further improving the predictive ability of the models; In the future, we will optimize the model through focusing on addressing sparse attention weight issues and refining sampling mechanisms to improve the accuracy and further balance the speed and accuracy of the proposed MIT-FRNet method.

CRediT authorship contribution statement

Jiayao Li: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Saihua Cai:** Resources, Writing – review & editing. **Li Li:** Data curation, Software. **Ruizhi Sun:** Supervision, Funding acquisition, Project administration. **Gang Yuan:** Investigation. **Rui Zhu:** Validation.

Declaration of competing 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.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

This work was partly supported by the National Key Research and Development Program of China (Grant no. 2021YFD1300101), the National Natural Science Foundation of China (NSFC) (Grant no. 62202206).

References

Araque, O., Zhu, G., & Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems, 165*, 346–359.

Bai, Z., Chen, X., Zhou, M., Yi, T., & Chien, W. C. (2021). Low-rank multimodal fusion algorithm based on context modeling. *Journal of Internet Technology, 22*(4), 913–921.

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning* (pp. 37–49). JMLR Workshop and Conference Proceedings.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (pp. 59–66). IEEE.

Cai, L., Wang, Z., Gao, H., Shen, D., & Ji, S. (2018). Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1158-1166). Association for Computing Machinery.

Chauhan, D. S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2019). Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5647-5657). Association for Computational Linguistics.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 960–964). IEEE.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.

Guo, W., Wang, J., & Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access, 7*, 63373–63394.

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* (pp. 1122-1131). Association for Computing Machinery.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Leidal, K., Harwath, D., & Glass, J. (2017). Learning modality-invariant representations for speech and images. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 424–429). IEEE.

Li, P., Laghari, A. A., Rashid, M., Gao, J., Gadekallu, T. R., Javed, A. R., & Yin, S. (2022). A deep multimodal adversarial cycle-consistent network for smart enterprise system. *IEEE Transactions on Industrial Informatics, 19*(1), 693–702.

Liu, R., Zhao, Y., Wei, S., Zheng, L., & Yang, Y. (2019). Modality-invariant image-text embedding for image-sentence matching. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15*(1), 1–19.

Liu, Z., Zhou, B., Chu, D., Sun, Y., & Meng, L. (2024). Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion, 101*, Article 101973.

Ma, F., Huang, S. L., & Zhang, L. (2021). An efficient approach for audio-visual emotion recognition with missing labels and missing modalities. In *2021 IEEE international conference on multimedia and Expo (ICME)* (pp. 1–6). IEEE.

Matsuura, T., Saito, K., Ushiku, Y., & Harada, T. (2018). Generalized bayesian canonical correlation analysis with missing modalities. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (pp. 641–656). Springer, Cham. https://doi.org/10.1007/978-3-030-11024-6_48.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (pp. 18-25).

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proceedings of the AAAI conference on artificial intelligence*, 34(02), 1359-1367. https://doi.org/10.1609/aaai.v34i02.5492.

Peng, M., Zhang, Q., & Huang, X. (2019). Weighed domain-invariant representation learning for cross-domain sentiment analysis. arXiv preprint arXiv:1909.08167.

Pham, H., Liang, P. P., Manzini, T., Morency, L. P., & Póczos, B. (2019). 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.v33i01.3301 6892

Qian, S., & Wang, C. (2023). COM: Contrastive masked-attention model for incomplete multimodal learning. *Neural Networks, 162*, 443–455.

Shang, C., Palmer, A., Sun, J., Chen, K. S., Lu, J., & Bi, J. (2017). VIGAN: Missing view imputation with generative adversarial networks. In *2017 IEEE International conference on big data (Big Data)* (pp. 766–775). IEEE.

Sun, Z., Sarma, P., Sethares, W., & Liang, Y. (2020). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 8992-8999. https://doi.org/10.1609/aaai.v34i05.6431.

Sun, H., Liu, J., Chen, Y. W., & Lin, L. (2023). Modality-invariant temporal representation learning for multimodal sentiment classification. *Information Fusion, 91*, 504–514.

Thongtan, T., & Phienthrakul, T. (2019). Sentiment classification using document embeddings trained with cosine similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 407-414). Association for Computational Linguistics.

Tran, L., Liu, X., Zhou, J., & Jin, R. (2017). Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1405-1414).

Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting* (pp. 6558–6569). NIH Public Access.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096-1103). Association for Computing Machinery.

Wang, Q., Ding, Z., Tao, Z., Gao, Q., & Fu, Y. (2021). Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing, 30*, 1771–1783.

Wei, S., Luo, Y., Ma, X., Ren, P., & Luo, C. (2023). MSH-Net: Modality-Shared Hallucination with Joint Adaptation Distillation for Remote Sensing Image Classification Using Missing Modalities. IEEE Transactions on Geoscience and Remote Sensing, 61, Article 4402615. https://doi.org/10.1109/TGRS.2023.3265650, URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10097714.

Xu, G., Meng, Y., Qiu, X., Yu, Z., & Wu, X. (2019). Sentiment analysis of comment texts based on BiLSTM. *IEEE Access, 7*, 51522–51532.

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J. & Yang, K. (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* (pp. 3718–3727). Association for Computational Linguistics.

Yuan, Z., Li, W., Xu, H., & Yu, W. (2021). Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 4400–4407), Association for Computing Machinery.

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250.

Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems, 31*(6), 82–88.

Zellinger, W., Moser, B. A., Grubinger, T., Lughofer, E., Natschläger, T., & Saminger-Platz, S. (2019). Robust unsupervised domain adaptation for neural networks via moment alignment. *Information Sciences, 483*, 174–191.

Zeng, J., Zhou, J., & Liu, T. (2022). Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. *IEEE Transactions on Multimedia, 25*, 6301–6314.

Zhang, C., Cui, Y., Han, Z., Zhou, J. T., Fu, H., & Hu, Q. (2020). Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence, 44*(5), 2402–2415.

Zhang, Y., Shen, J., Zhang, Z., & Wang, C. (2021). Partial Modal Conditioned GANs for Multi-modal Multi-label Learning with Arbitrary Modal-Missing. In *Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part II 26* (pp. 413-428). Springer International Publishing.

Zhang, R., Wang, C., & Liu, C. L. (2023). Cycle-consistent weakly supervised visual grounding with individual and contextual representations. *IEEE Transactions on Image Processing, 32*, 5167–5180.

Zhao, J., Li, R., & Jin, Q. (2021). Missing modality imagination network for emotion recognition with uncertain missing modalities. In *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)* (pp. 2608-2618). Online. Association for Computational Linguistics.

Zhou, T., Ruan, S., & Hu, H. (2022). A literature survey of MR-based brain tumor segmentation with missing modalities. *Computerized Medical Imaging and Graphics, 104*, Article 102167. https://doi.org/10.1016/j.compmedimag.2022.102167