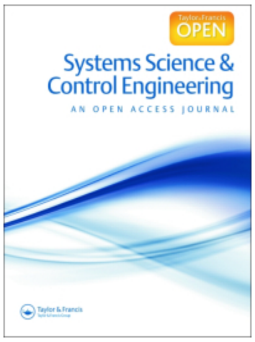


Systems Science & Control Engineering

An Open Access Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tssc20

A systematic literature review on incomplete multimodal learning: techniques and challenges

Yifan Zhan, Rui Yang, Junxian You, Mengjie Huang, Weibo Liu & Xiaohui Liu

To cite this article: Yifan Zhan, Rui Yang, Junxian You, Mengjie Huang, Weibo Liu & Xiaohui Liu (2025) A systematic literature review on incomplete multimodal learning: techniques and challenges, Systems Science & Control Engineering, 13:1, 2467083, DOI: 10.1080/21642583.2025.2467083

To link to this article: https://doi.org/10.1080/21642583.2025.2467083

Published online: 26 Feb 2025.

SYSTEMS SCIENCE & CONTROL ENGINEERING: AN OPEN ACCESS JOURNAL
2025, VOL. 13, NO. 1, 2467083
https://doi.org/10.1080/21642583.2025.2467083

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

A systematic literature review on incomplete multimodal learning: techniques and challenges

Yifan Zhan[a,b], Rui Yang[a], Junxian You[a,b], Mengjie Huang[c], Weibo Liu[d] and Xiaohui Liu[d]

[a]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, People's Republic of China; [b]School of Electrical Engineering, Electronics and Computer Science, University of Liverpool, Liverpool, UK; [c]Design School, Xi'an Jiaotong-Liverpool University, Suzhou, People's Republic of China; [d]Department of Computer Science, Brunel University London, Uxbridge, UK

ABSTRACT

Recently, machine learning technologies have been successfully applied across various fields. However, most existing machine learning models rely on unimodal data for information inference, which hinders their ability to generalize to complex application scenarios. This limitation has resulted in the development of multimodal learning, a field that integrates information from different modalities to enhance models' capabilities. However, data often suffers from missing or incomplete modalities in practical applications. This necessitates that models maintain robustness and effectively infer complete information in the presence of missing modalities. The emerging research direction of incomplete multimodal learning (IML) aims to facilitate effective learning from incomplete multimodal training sets, ensuring that models can dynamically and robustly address new instances with arbitrary missing modalities during the testing phase. This paper offers a comprehensive review of methods based on IML. It categorizes existing approaches based on their information sources into two main types: based on internal information and external information methods. These categories are further subdivided into data-based, feature-based, knowledge transfer-based, graph knowledge enhancement-based, and human-in-the-loop-based methods. The paper conducts comparative analyses from two perspectives: comparisons among similar methods and comparisons among different types of methods. Finally, it offers insights into the research trends in IML.

ARTICLE HISTORY
Received 19 November 2024
Accepted 10 February 2025

KEYWORDS
Incomplete multimodal learning; multimodal learning; modality missing

1. Introduction

Conventional deep learning methods mainly rely on unimodal data for feature extraction and information inference. These models are primarily trained on data from a single modality, achieving prediction or classification by analyzing features within that specific modality and applying them to the unimodal dataset (Z. Chen, Yang, Huang, Li, et al., 2024; Ghandi et al., 2023; X. Qian & Cui, 2023; Qin et al., 2024). In image processing (Azad et al., 2024; S. Zhang & Metaxas, 2024), the image exists in the form of pixel points and has a high degree of spatial structure. Convolutional Neural Networks (CNNs) effectively extract edge, shape, and texture features from images through mechanisms like local receptive fields and parameter sharing, enabling them to excel in image-related tasks (Diwan et al., 2023; Joseph et al., 2021; X. Li et al., 2023; Y. Liu, Sun, et al., 2021; Xiao et al., 2020). In natural language processing (NLP) (Nadkarni et al., 2011; Kang et al., 2020; Lauriola et al., 2022), text data is usually represented as discrete vocabulary with sequential properties. Deep learning models, like Recurrent Neural Networks (RNNs) (Yin et al., 2017) and Long Short-Term Memory (LSTM) (Y. Yu et al., 2019), are capable of modelling sequential dependencies within the text, resulting in a good performance in NLP tasks (Min et al., 2023). Similarly, speech data also exhibits temporal features. RNNs and their variants can capture the temporal correlation of speech signals, leading to achievement in tasks such as speech recognition (Y. Yu et al., 2019).

Despite the substantial advancements achieved by unimodal deep learning models, unimodal data often fails to provide complete information in complex applications, leading to inaccurate or biased judgments (Dong et al., 2023; F. Han et al., 2023; Qin, Yang, et al., 2023). In image classification, if models only infer according to the visual information of the image and neglect the textual descriptions, they may lead to reduced accuracy and decreased generalization (L. Chen et al., 2021; Kim et al., 2022). In speech recognition tasks, while the speech signal provides audio features of the language, the model's recognition accuracy often struggles to improve further without additional information,

CONTACT Rui Yang R.Yang@xjtlu.edu.cn

such as facial expressions in a video or accompanying text (Alharbi et al., 2021; Oruh et al., 2022; Ravanelli et al., 2020). To overcome these challenges, multimodal learning has gradually drawn attention for its capability to combine information of different modalities (Bayoudh et al., 2022; P. Xu et al., 2023).

Compared to unimodal deep learning, multimodal learning provides a more comprehensive perspective by simultaneously utilizing multimodal data, thereby capturing complex relationships within the data more effectively (J. Han et al., 2019; B. Li et al., 2020; P. P. Liang et al., 2024; K. Sharma & Giannakos, 2020). For example, in image captioning tasks, the model not only needs to analyze the content of the image but also to understand the related textual information to generate accurate and contextually relevant descriptions (Hossain et al., 2019; Stefanini et al., 2022). In sentiment analysis, multimodal learning can combine text, audio, and video information to assess the user's emotional state, remarkably improving the accuracy of the analysis (Z. Liu et al., 2024; Y. Sun et al., 2024). Furthermore, in medical diagnosis, by integrating medical images with clinical text records, models can evaluate patients' health conditions comprehensively, assisting doctors in making precise diagnoses (Azad et al., 2022; Qiu et al., 2024). Recently, self-attention mechanisms and transformer models have been widely utilized, enhancing the ability to process multimodal data, particularly in capturing the complex relationships between modalities (S. Qian & Wang, 2023; Y. Zhan & Yang, 2023). Additionally, emerging technologies such as cross-modal adversarial training and transfer learning have been introduced to strengthen the model's generalization across different modalities (Ben-Cohen et al., 2019; X. Chen et al., 2023; M. Li et al., 2022; H. Wang, Ma, et al., 2024).

Most existing multimodal learning-based models operate under the assumption that data from all modalities is always available, a premise often challenging to achieve in practical applications. Considerations such as human error, privacy issues, and ambient changes frequently give rise to modality missing or incomplete multimodal data (Cai et al., 2018; K. Sharma & Giannakos, 2020; Y. Shen & Gao, 2019). Modality missing refers to the situation where the missing or unavailability of certain modalities causes a decline in model effectiveness (S. Yu et al., 2024). For autonomous driving, LiDAR might lose data because of object occlusion or adverse conditions (Roche et al., 2021; Zheng et al., 2023). In power grid monitoring, faults or communication issues can disrupt real-time data transmission, while incorrect patient positioning or equipment malfunctions can degrade medical images or cause data loss (G. Li et al., 2021). Moreover, ethical and privacy concerns may prevent the disclosure of sensitive data. When one or more modalities are unavailable, the model may lose critical contextual information, remarkably dropping its performance and potentially leading to task failure or critical errors in judgment (H. Liu, Wei, et al., 2023; Y. Shen & Gao, 2019).

Therefore, the study of Incomplete Multimodal Learning (IML) is significant. The challenge of IML is finding ways to effectively learn from incomplete multimodal training sets while ensuring that the model can dynamically and robustly handle new instances with any missing modalities during the testing phase (S. Qian & Wang, 2023; Y. Wang et al., 2024). To solve this problem, researchers have proposed various strategies, including data generation-based, shared feature-based, and knowledge transfer-based methods (Lee et al., 2023; M. Li et al., 2022; Matsuura et al., 2018; Zeng, Liu, et al., 2022). Data generation-based methods enhance model completeness and performance by generating data for missing modalities (Islam et al., 2021). Techniques such as Generative Adversarial Networks (GANs) use a game-theoretic approach to generate synthetic data, while Variational Autoencoders (VAEs) model the data distribution to create missing modalities (Q. Wang, Ding, et al., 2018; Y. Wang, Zhou, et al., 2018). Shared feature-based methods aim to create shared representations across different modalities, allowing the model to leverage features from other modalities even when some are absent (R. Liu et al., 2024; H. Wang et al., 2023). Knowledge transfer-based methods tackle the problem of missing modalities by transferring knowledge from complete modalities or related tasks (Ding et al., 2014, 2015). The differences among conventional deep learning-based, complete multimodal learning-based, and IML-based methods are shown in Figure 1.

Recently, some reviews have explored the issue of modality missing and its development in brain tumor segmentation (Azad et al., 2022; Zhou et al., 2023). However, comprehensive reviews of IML-based methods across various fields are still lacking. As research advances, new technologies and findings based on IML continue to emerge. Therefore, this paper reviews the progress of research based on IML, categorizes relevant literature according to the sources of information for addressing the issue, and provides a comprehensive analysis and discussion, along with suggestions for future directions. The contribution of this review lies in filling the current gaps in the existing research, outlined as follows:

(1) This review systematically outlines the development of IML-based methods across various fields and

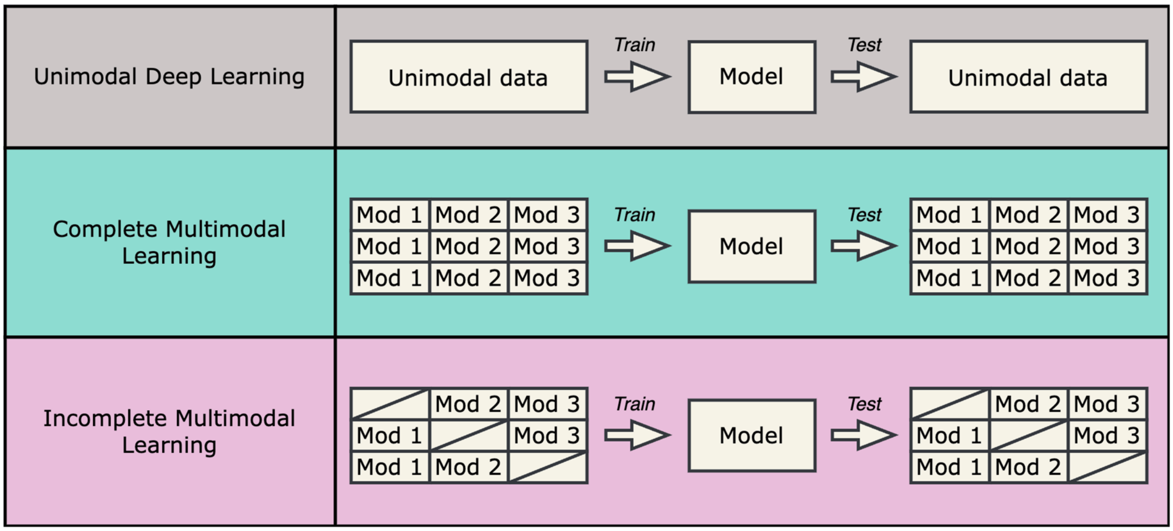


Figure 1. Differences among conventional deep learning-based, complete multimodal learning-based, and IML-based methods (using three modalities as an example, where 'mod' represents a modality and the box with diagonal lines indicates that a modality is missing for the sample).

provides a structured guideline for researchers to address the challenges of modality missing. By clarifying the core goal of IML and offering a relevant definition of the modality missing problem, this work allows readers to gain deep insights and develop effective strategies to tackle the modality missing problem;

(2) This review proposes a novel taxonomy for IML from the perspective of the information used to address modality missing and provides a comprehensive overview of IML-based methods for each type of category, aiming to help researchers and practitioners efficiently select suitable models;

(3) This paper provides a comparative analysis of IML-based methods from the perspectives of both similarities and differences, analyzes their strengths and weaknesses, and identifies future research directions.

The subsequent sections of this review are arranged as follows: Section 2 presents the problem formulation and related definitions of IML; Section 3 provides a detailed introduction of methods based on IML; Section 4 qualitatively analyzes these methods from the perspective of similar and different methods; Section 5 discusses the challenges and opportunities in existing methods; Finally, Section 6 presents the conclusion.

2. Preliminary

2.1. Problem formulation of modality missing

Considering a dataset D containing M modalities and N multimodal samples, the formulation of the missing modality is introduced. For $m \in \{1, 2, \ldots, M\}$, the representation of the mth modality for the nth multimodal sample is defined as $x_n^m \in \mathbb{R}^{d_m}$, with d_m being the dimension of the mth modality. For one sample, there are $2^M - 1$ possible cases of modality missing (excluding the case where all modalities of one sample are missing). An indicator matrix $J \in \mathbb{R}^{M \times N}$ for multimodal samples is denoted as:

$$J_n^m = \begin{cases} 1, & \text{if the } n\text{th sample has the } m\text{th modality} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where each column of J represents the modality status of the corresponding sample. For each incomplete sample x_n^m, there exists $\sum_{m=1}^{M} J_n^m < M$. IML aims to find ways to effectively learn from incomplete multimodal training sets while ensuring that the model can dynamically and robustly handle new instances with any missing modalities during the testing phase.

2.2. Assessing incompleteness in multimodal datasets

An improved method is proposed in S. Yu et al. (2024) for comprehensively assessing modality incompleteness in multimodal datasets. Conventional methods typically use a modality missingness score η to describe the extent of missing data across modalities in a dataset. Specifically, η is calculated as follows:

$$\eta = 1 - \frac{1}{N \times M} \sum_{n=1}^{N} \sum_{m=1}^{M} J_n^m \tag{2}$$

where $\sum_{n=1}^{N}\sum_{m=1}^{M} J_n^m$ denotes the total number of existing modalities.

The above calculation overlooks whether the missingness is balanced across modalities. Even if two datasets share the same η, one may exhibit significantly greater incompleteness if one modality is substantially underrepresented. This imbalance leads to asymmetry in the available information, posing challenges for a model in extracting consistent representations from each modality. To address this issue, a modality imbalance coefficient β is introduced to calculate as follows:

$$\beta = \exp\left(-\mathrm{std}\left(\sum_{n=1}^{N} J_n^1, \sum_{n=1}^{N} J_n^2, \ldots, \sum_{n=1}^{N} J_n^M\right)\right) \tag{3}$$

where $\mathrm{std}(\cdot)$ is the standard deviation function measuring the distributional disparity of sample counts across modalities. A higher standard deviation indicates a greater imbalance between modalities, resulting in a lower β. Finally, a composite metric ξ is utilized to present modality incompleteness:

$$\xi = \left(\eta + \frac{1}{1+\exp(\beta)}\right)/2 \tag{4}$$

Therefore, incompleteness is measured by both the degree of modality missingness and the degree of modality imbalance.

2.3. Related definitions

(1) *Modality*: modality refers to the way of expressing or perceiving things and is the specific manifestation of a source or form of information. For example, the human sensory system includes touch, hearing, vision, and smell, each corresponding to a different modality. Additionally, information media can also be considered different modalities, such as speech, video, text, etc. Various sensor data, such as radar, infrared, accelerometers, and others, can likewise be regarded as different modalities. When a research problem or dataset includes multiple such modalities of information, it is referred to as a multimodal problem (Bayoudh et al., 2022; M. Ma et al., 2021; P. Xu et al., 2023).

(2) *Multimodal learning*: multimodal learning is a method that utilizes data from multiple modalities (or signals) simultaneously for learning and reasoning.

(3) *Homogeneous modality*: homogeneous modality refers to modalities with the same data type and characteristics, such as images captured from different angles or texts in different languages (W. Liang et al., 2021; Ye et al., 2020). Although their content may differ, their data characteristics are similar. In multimodal learning with homogeneous modalities, the model typically needs to integrate and compare similar data to gain information from different perspectives.

(4) *Heterogeneous modality*: heterogeneous modality refers to modalities with entirely different data types and characteristics, such as images and texts, or audio and video. These modalities have significantly different data characteristics and forms of representation (J. Chen & Zhang, 2020; Suzuki & Matsuo, 2022; Z. Zhang et al., 2022). In multimodal learning with heterogeneous modalities, the model is required to handle the heterogeneity between different modalities and effectively integrate them to utilize the complementary information from each modality (Bayoudh et al., 2022; M. Ma et al., 2021; P. Xu et al., 2023).

3. Incomplete multimodal learning-based methods

This section introduces IML-based methods for addressing the issue of modality missing. As shown in Figure 2, methods are classified into two types according to the source of information for addressing the issue: based on internal information and external information methods. Methods based on internal information rely on data and patterns from the existing dataset to handle missing modalities (Q. Wang, Ding, et al., 2018, 2020; R. Wu et al., 2020; C. Zhang et al., 2020). In contrast, methods based on external information rely on external resources beyond the dataset, such as additional context, additional datasets, domain knowledge or human involvement (Buciluă et al., 2006; Hinton, 2015; D. Hu et al., 2019; Ou et al., 2023; Qin, Zhang, et al., 2023; Zanzotto, 2019). For a comprehensive overview, Table 1 provides a summary of representative IML-based studies.

3.1. Internal information-based methods

Methods based on internal information are classified into data-based and feature-based approaches according to the level at which they handle missing modalities. Data-based methods focus on completing or grouping missing modalities before the data is input into the downstream task model (Cai et al., 2018; Shang et al., 2017; Y. Wang et al., 2024; H. Yang et al., 2023). Feature-based methods focus on the feature level by mapping different modalities into a latent space to get shared information (P. Li et al., 2022; Q. Wang, Ding, et al., 2020; R. Zhang et al., 2023; Zhao et al., 2021).

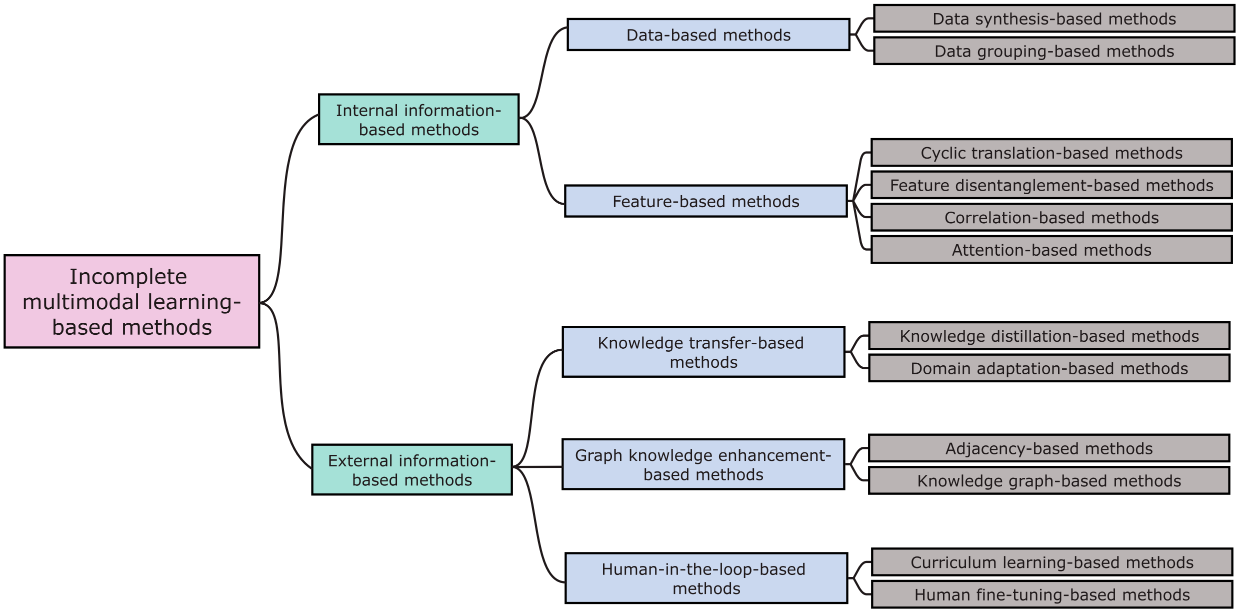

Incomplete multimodal learning-based methods
Internal information-based methods
External information-based methods
Data-based methods
Feature-based methods
Knowledge transfer-based methods
Graph knowledge enhancement-based methods
Human-in-the-loop-based methods
Data synthesis-based methods
Data grouping-based methods
Cyclic translation-based methods
Feature disentanglement-based methods
Correlation-based methods
Attention-based methods
Knowledge distillation-based methods
Domain adaptation-based methods
Adjacency-based methods
Knowledge graph-based methods
Curriculum learning-based methods
Human fine-tuning-based methods

Figure 2. Taxonomy of IML-based methods.

Table 1. Summary of representative IML-based literature. 'Homo' represents homogeneous, while 'Hetero' denotes heterogeneous.

Reference	Application Scenarios	Taxonomy	Modality	Number of Modalities	Published Date
Islam et al. (2021)	Brain tumor segmentation	Data synthesis-based	Homo	4	2021
Zhou et al. (2020)	Medical image synthesis	Data synthesis-based	Homo	3	2020
Y. Zhang et al. (2024)	Medical image synthesis	Data synthesis-based	Homo	4	2024
Y. Zhang et al. (2024)	Medical image synthesis	Data synthesis-based	Homo	4	2024
M. Ma et al. (2021)	Multimodal generation	Data synthesis-based	Hetero	2/3	2021
Y. Wang et al. (2024)	Emotion recognition	Data synthesis-based	Hetero	3	2024
Y. Sun et al. (2024)	Sentiment analysis	Data synthesis-based	Hetero	3	2024
Yuan et al. (2012)	Disease prediction	Data grouping-based	Hetero	2/3	2012
Xiang et al. (2013)	Disease prediction	Data grouping-based	Hetero	4	2013
Zhao et al. (2021)	Emotion recognition	Cyclic translation-based	Hetero	3	2021
Q. Yang et al. (2022)	Brain tumor segmentation	Feature disentanglement-based	Homo	4	2022
R. Liu et al. (2024)	Emotion recognition	Feature disentanglement-based	Hetero	3	2024
S. Yu et al. (2024)	Federated learning	Correlation-based	Homo	2/3	2024
Qiu et al. (2024)	Brain tumor segmentation	Correlation-based	Homo	4	2024
G. Yang et al. (2024)	Fault diagnosis	Correlation-based	Homo	8	2024
J. Shi et al. (2023)	Brain tumor segmentation	Attention-based	Homo	4	2023
Lee et al. (2023)	Visual recognition	Attention-based	Hetero	2	2023
Zeng, Zhou, et al. (2022b)	Sentiment analysis	Attention-based	Hetero	3	2022
S. Qian and Wang (2023)	Multimodal learning	Attention-based	Hetero	2/3	2023
J. Li et al. (2024)	Emotion recognition	Attention-based	Hetero	3	2024
Maheshwari et al. (2024)	Semantic segmentation	Knowledge distillation-based	Homo	2	2024
C. Chen et al. (2021)	Brain tumor segmentation	Knowledge distillation-based	Homo	3/4	2021
H. Liu, Wei, et al. (2023)	Brain tumor segmentation	Knowledge distillation-based	Homo	4	2023
D. Zhang et al. (2024)	Brain tumor segmentation	Knowledge distillation-based	Homo	4	2024
W. Zhang et al. (2021)	Visual recognition	Domain adaptation-based	Homo	2	2021
Y. Shen and Gao (2019)	Brain tumor segmentation	Domain adaptation-based	Homo	4	2019
Malitesta et al. (2024)	Recommender system	Adjacency-based	Hetero	2	2024
C. Zhang et al. (2022)	Healthcare informatics	Adjacency-based	Hetero	4	2022
Y. Liang (2024)	Recommender system	Knowledge graph-based	Hetero	2	2024
X. Lu et al. (2022)	Representation learning	Knowledge graph-based	Hetero	3	2022
Z. Chen et al. (2023)	Entity alignment	Knowledge graph-based	Hetero	4	2023
A. Sharma and Hamarneh (2019)	Medical image synthesis	Curriculum learning-based	Homo	4	2019
D. Hu et al. (2019)	Sensory substitution	Human fine-tuning-based	Hetero	2	2019
Ou et al. (2023)	Remote sensing image	Human fine-tuning-based	Hetero	2	2023

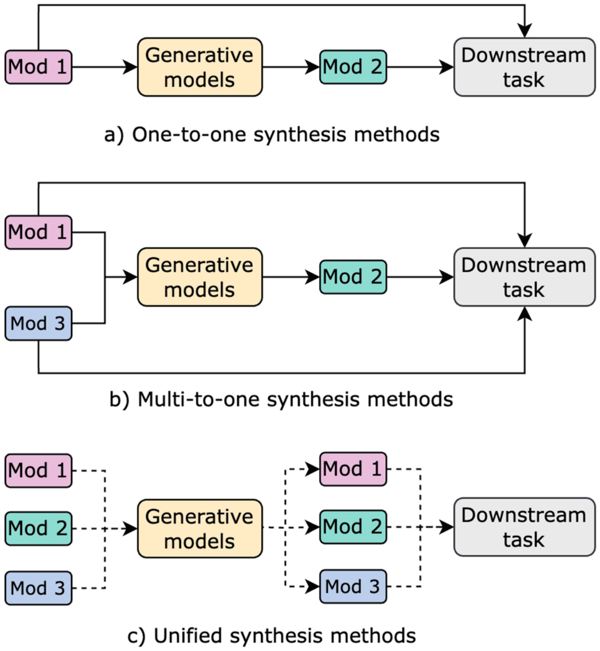
3.1.1. Data-based methods

These methods aim to mitigate the impact of missing modalities at the data level and are categorized into data synthesis-based and data grouping-based methods.

3.1.1.1. Data synthesis-based methods. To tackle modality missing, the most direct solution is to synthesize missing data. Conventional solutions are primarily reliant on data imputation, such as zero imputation (Zheng et al., 2023) and mean imputation (Y. Sun et al., 2024; C. Zhang et al., 2020), but these methods have limited effectiveness as they fail to accurately model the complex relationships and underlying structures of the data (Van Tulder & de Bruijne, 2015). Recently, data synthesis techniques have gradually replaced imputation methods (Shang et al., 2017). Data synthesis-based methods seek to use generative models to provide complete modalities for downstream tasks (Hao et al., 2024). Specifically, these methods refer to first inferring the missing modalities using neural network-based generative models to gain a complete multimodal dataset, and applying state-of-the-art multimodal learning models for downstream tasks (Dalmaz et al., 2022).

Data synthesis-based methods primarily employ generative networks like VAEs (Kingma, 2013), GANs (Goodfellow et al., 2014), and diffusion models (Ho et al., 2020) to generate missing modality data. VAEs-based methods (M. Ma et al., 2021; M. Shen et al., 2021; Y. Shi et al., 2019; Sutter et al., 2020; Tran et al., 2017; M. Wu & Goodman, 2018) map input data to a latent space and samples from it to generate the missing modalities, balancing reconstruction loss and regularization of the latent space to generate stable and diverse modality data. In GANs-based approaches for handling missing modalities (Cai et al., 2018; Pan et al., 2021; Shang et al., 2017; Suo et al., 2019; Q. Wang, Ding, et al., 2018, 2020; R. Wu et al., 2020; C. Zhang et al., 2020), the generator produces the missing modality data, while the discriminator differentiates between generated and real data. Through adversarial training, the generator is optimized to produce realistic missing modalities for downstream tasks (Cai et al., 2018). Recently, diffusion models-based methods have generated missing modalities by gradually adding noise and denoising, producing high-quality data and reducing semantic ambiguity (Croitoru et al., 2023; Y. Wang et al., 2024).

Based on the number of observable and target modalities, data synthesis methods can be categorized into three types: one-to-one synthesis methods, multi-to-one synthesis methods, and unified synthesis methods (D. Zhang et al., 2024). An overview of these three strategies is illustrated in Figure 3. One-to-one synthesis refers to generating a target modality from a complete observed modality. For example, cross-modal medical image synthesis was achieved on limited paired data and abundant unpaired data through joint dictionary learning with weak coupling and geometric co-regularization, leading to improved synthesis quality using sparse representation (Huang et al., 2017). To tackle differences between modalities, VIGAN (Shang et al., 2017) treats each view as an independent domain, utilizes GANs for inter-domain mapping, and uses a multimodal denoising autoencoder (DAE) to reconstruct the missing views from outputs of GANs. Similarly, GANs (Cai et al., 2018) were used to generate missing modalities, and the quality of the generated modalities was enhanced through content loss and adversarial loss, even in the absence of class labels. However, the one-to-one synthesis methods necessitate training a distinct model for each missing condition. This requirement significantly escalates the computational burden when attempting to synthesize multiple modalities.



a) One-to-one synthesis methods

b) Multi-to-one synthesis methods

c) Unified synthesis methods

Figure 3. Basic framework of three types of data synthesis methods. The dashed arrows in the figure represent input pathways for arbitrary modality combinations.

The multi-to-one synthesis method aims to use multiple available modalities to get a single target modality. Compared to one-to-one methods, this approach better captures the shared features between modalities, optimizing the synthesized quality (J. Liu et al., 2023). However, it faces the challenge of effectively extracting and integrating features from multiple modalities to avoid information loss or feature conflicts. Therefore, a synthesis model was introduced (Islam et al., 2021), built upon Fully Convolutional Networks (FCN) (Long et al., 2015) and Conditional GANs, to reduce feature redundancy and model sparsity. Besidesa multi-scale gate mergence was introduced to automatically learn weights for various modalities (B. Zhan et al., 2021), improving task-related information while minimizing irrelevant information. Furthermore, a layer-wise fusion strategy was employed to combine multimodal representations adaptively (Zhou et al., 2020). In addition, the incomplete multimodality-diffused emotion recognition (IMDer) was introduced (Y.

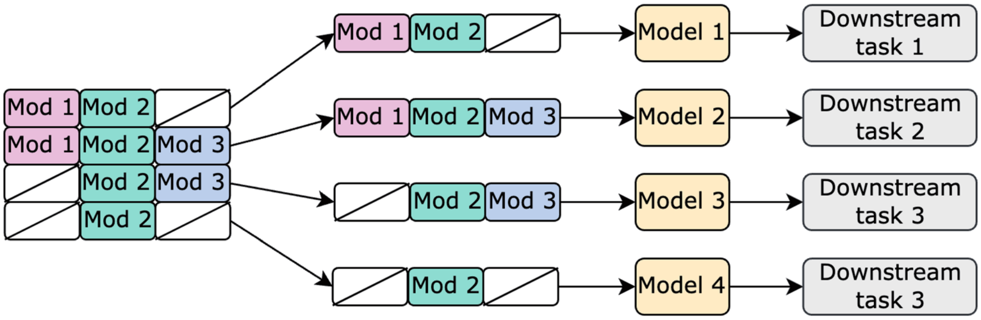


Figure 4. Basic framework of data grouping-based methods.

Wang et al., 2024), using a score-based diffusion model to reconstruct missing data by ensuring distribution consistency and semantic alignment. Although the multi-to-one synthesis method effectively captured shared features, it struggled to address multi-target modality generation tasks.

The unified synthesis method refers to generating one or more target modalities from any available modalities in a single forward pass. The main challenge of this approach lies in designing an efficient network structure that can handle various missing modality scenarios while ensuring high quality and consistency in the generated modalities. In response to the challenge of severe modality absence, Bayesian meta-learning was leveraged to predict prior weights through a reconstruction network, effectively addressing significant data loss in multimodal learning (M. Ma et al., 2021). Additionally, a GANs-based approach with a decoupling scheme (L. Shen et al., 2020) was introduced to extract shared content encoding and independent style encoding. More recently, a unified adaptive multimodal image synthesis method was proposed, using a shared super-encoder to embed the features of each modality into a shared space, followed by a graph attention fusion module to ensure consistency in the generated results (H. Yang et al., 2023). Furthermore, Transformer was employed to model long-range dependencies between different modalities, achieving unified image synthesis (Dalmaz et al., 2022). Finally, the Commonality- and Discrepancy-Sensitive Encoder (CDS-Encoder) was proposed, and a Dynamic Feature Unification Module (DFUM) was introduced to integrate information from different combinations of available modalities (Y. Sun et al., 2024; Y. Zhang et al., 2024).

3.1.1.2. Data grouping-based methods.

These methods address incomplete data samples by partitioning them according to available data sources and training separate models for each group, transforming the problem into a multi-task learning scenario (S. Qian & Wang, 2023). An overview of this strategy is depicted in Figure 4.

Two methods for incomplete data fusion were designed (Yuan et al., 2012). The first, the incomplete multi-source feature learning (iMSF) framework, divided the data into blocks based on missing modalities and built separate models for each block, enhancing model sparsity and performance through joint feature learning and regularization. The second method, the model score completion scheme (ScoreComp), involved training base models on each data source to create a prediction score matrix, which was then completed using imputation methods to form a full score matrix for training the final classifier.

Unlike iMSF, the data groups in Incomplete Source-Feature Selection (iSFS) model overlap to make each group contain more samples (Xiang et al., 2013), as illustrated in Figure 5, where each red box indicates a group. The model was trained separately on each group of data sources, applying different regularization strategies to each data source by adjusting regularization parameters. This approach performed both feature-level and source-level analysis simultaneously and simplified the optimization problem through equivalence transformation, ultimately constructing a multi-source fusion model.

Common issues associated with data grouping-based methods include sample imbalances among subsets, which can cause the model to favor larger groups during training and impact generalization (J. Sun et al., 2024). Additionally, defining appropriate grouping criteria for complex multimodal datasets is challenging and may overlook global feature dependencies, hindering effective information integration.

3.1.2. Feature-based methods

Unlike data-based methods, feature-based methods handle missing modality data at the feature level by projecting data from each modality into a latent space to identify shared information across modalities (Azad et al., 2022; Zhou et al., 2023). These methods typically adopt an end-to-end multimodal learning framework, where the handling of missing modalities and the

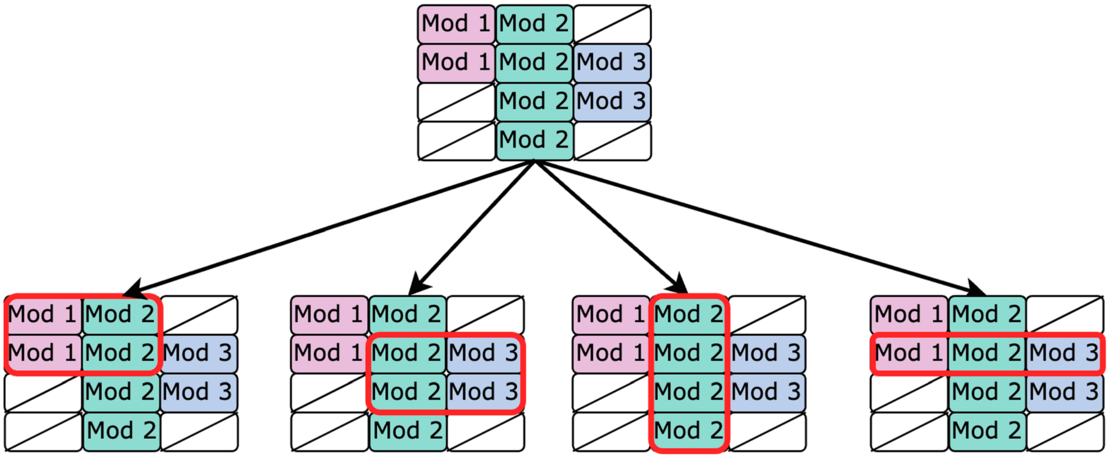


Figure 5. Illustration of data grouping in iSFS.

downstream task are integrated within a unified model. Feature-based methods can be categorized into four types: cyclic translation-based, feature disentanglement-based, correlation-based, and attention-based methods.

3.1.2.1. Cyclic translation-based methods. Cycle consistency is an essential principle in multimodal learning, stating that when transforming between different modalities, the final result should be consistent with the original input (W. Sun et al., 2021; Zhao et al., 2021). Cycle translation is a method for achieving cycle consistency, ensuring that during modality transformations – such as converting an image to text and then back to an image – the final output is similar to the original input (C.-T. Lin et al., 2020).

Some data synthesis-based methods employed cycle consistency to generate data. For instance, a generative partial multi-view clustering model was introduced (Q. Wang, Ding, et al., 2020), which used adaptive fusion and cycle consistency to generate missing views from shared representations of other views. However, such methods increase model complexity and computational cost, and may face risks of error accumulation and limited generalization ability.

To address these issues, cyclic translation-based methods use cycle consistency to learn joint representations across modalities and perform downstream tasks, as depicted in Figure 6. The Multimodal Cyclic Translation Network (MCTN) was proposed (Pham et al., 2019), a Seq2Seq-based model that learned joint features through cyclic transformations between modalities, using mean squared error for cycle consistency. To handle three modalities, MCTN Trimodal employed two Seq2Seq models for cyclic translation, allowing multimodal representations from a single input during testing. However, experiments showed training asymmetry, with results differing based on the choice of source and target modalities.

Similarly, a Deep Multimodal Adversarial Cycle-Consistent Network (DMACCN) was proposed (P. Li et al., 2022), which captured intrinsic data patterns by modelling local structures and global topology. An adversarial cycle-consistent loss in DMACCN effectively guided clustering and cross-modal semantic alignment while fusing complementary information and capturing clustering structures. Furthermore, cycle consistency constraints were introduced to strengthen related region-phrase pairs and weaken unrelated ones (R. Zhang et al., 2023), leveraging bidirectional associations to reduce matching ambiguities. However, the aforementioned models are designed for the case of fixed modality missing and face challenges in generalizing to randomly missing modalities.

To address randomly missing modalities, the Missing Modality Imagination Network (MMIN) was proposed (Zhao et al., 2021), employing Cascade Residual Autoencoder (CRA) and cycle consistency learning to create robust joint multimodal representations. MMIN inferred the representation of the missing modality through forward imagination with available modalities and estimated the representation of the original modality through backward imagination. During testing, MMIN predicted representations for any missing modality condition.

3.1.2.2. Feature disentanglement-based methods. Feature disentanglement aims to decompose complex feature representations into independent and interpretable components (Azad et al., 2022; Q. Yang et al., 2022; Zhou et al., 2023). This process enables the model to identify the role of each feature in different contexts, avoiding interference between features and thereby improving generalization (Y. Chen et al., 2023). In tasks such as image generation, feature disentanglement helps the model

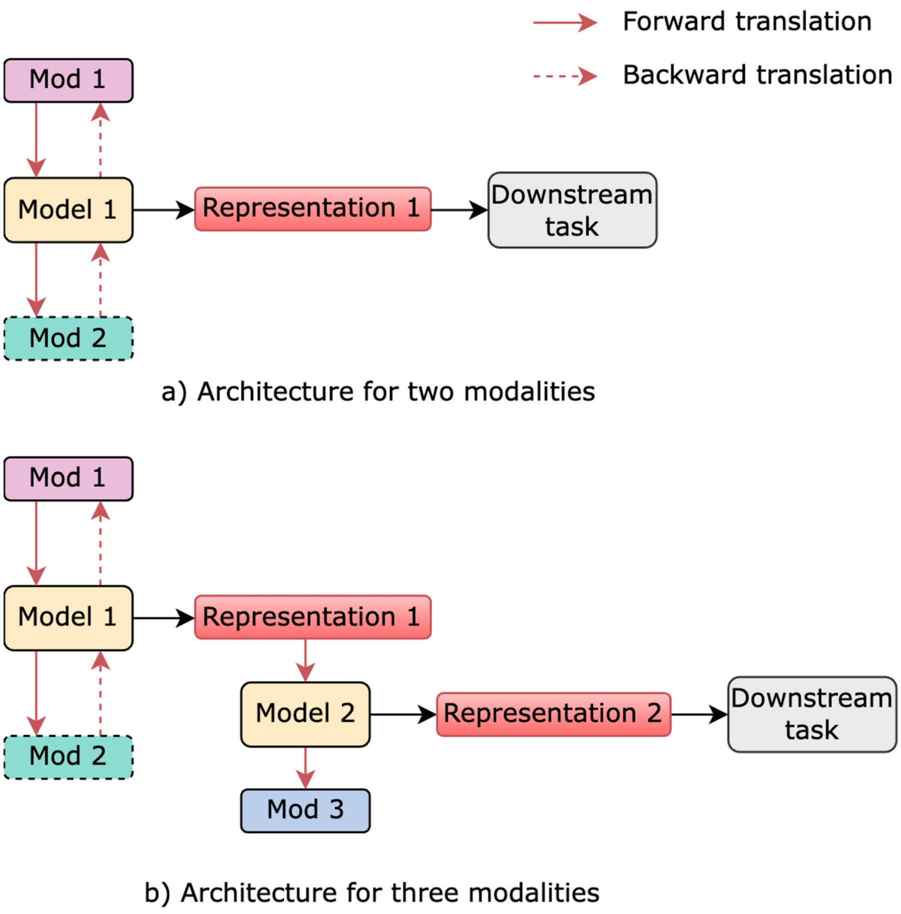


Figure 6. Two basic frameworks of cyclic translation-based methods, with dashed lines indicating the modality is incomplete.

separate features like pose, colour, and shape, resulting in accurate outputs (Q. Yang et al., 2022).

The core idea of feature disentanglement-based models is to construct comprehensive representations by extracting modality-specific and modality-shared representations from each modality, as shown in Figure 7. Specifically, the model establishes specific models for each modality to extract unique specific features. Additionally, a shared model is created for all modalities to obtain shared representations. Both the specific and the shared representations are then used together for downstream tasks. This approach allows the model to effectively address the issue of missing modalities while enhancing the integration and utilization of diverse information in multimodal learning.

A dual-decoupling network called D^2-Net was proposed for brain tumor segmentation with missing modalities (Q. Yang et al., 2022). This method decoupled modality-specific information through a spatial-frequency joint modality contrastive learning scheme and then guided the extraction of dense tumor region knowledge by aligning the decoupled binary teacher network features with the student network. Furthermore, the 'Disentangle First, Then Distill' (DFTD) framework was introduced for completing missing medical image modalities (Y. Chen et al., 2023). This framework first decoupled the image into cross-modality correlated and modality-specific representations using a region-aware disentanglement module. It then used an imputation-induced distillation module to fill in missing modality representations by utilizing inter-modality correlated features.

Similarly, modality-invariant features were leveraged to generate missing modality information (R. Liu et al., 2024). Pre-training was conducted using contrastive learning on complete modalities, and self-supervised learning was used to train an invariance encoder to extract modality-invariant features. The robust imagination module then reconstructed the missing information based on these features, combining the generated modality with the available modalities for emotion recognition.

Most feature disentanglement-based methods are typically designed for specific tasks (such as classification or segmentation) and primarily address modality missing during the evaluation phase (Y. Chen et al., 2023; R. Liu et al., 2024; Q. Yang et al., 2022). To address these issues,

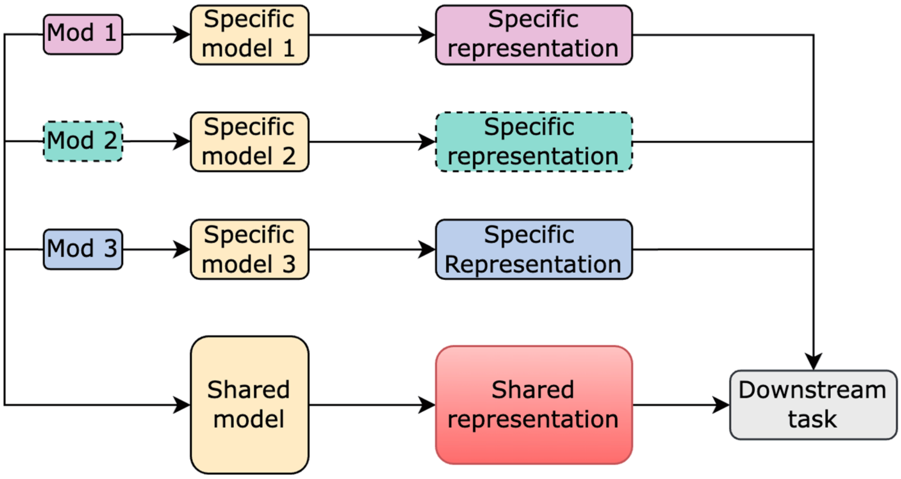


Figure 7. Basic framework of feature disentanglement-based methods, with dashed lines indicating the modality is incomplete.

the ShaSpec was introduced to generate features within a shared space by modelling both shared and specific features (H. Wang et al., 2023). This approach integrated auxiliary task strategies such as distribution alignment and domain classification, and employed a residual feature fusion mechanism. Therefore, the method effectively handled the modality missing during both the training and testing phases and was applicable to different tasks.

3.1.2.3. Correlation-based methods. Researchers have found that there are interconnections in information and features between different modalities and samples. This correlation is evident not only in the sharing of information and the complementarity of features but also in the consistent structures they follow within the same context (Azad et al., 2022; S. Qian & Wang, 2023; Zhou et al., 2023). Correlation-based methods model the relationships either between samples or between modalities, capturing their underlying correlations, as depicted in Figure 8.

A module based on deep canonical correlation analysis was employed to enhance feature alignment by maximizing correlation within matched categories during training (Q. Wang, Lian, et al., 2020).Moreover, a likelihood-based method was designed to characterize conditional distributions associated with samples of both complete and incomplete modalities (F. Ma et al., 2021). Additionally, constraints from the Hilbert-Schmidt Independence Criterion (HSIC) were applied to guide the model in completing missing features (Y. Liu, Fan, et al., 2021).

Besides, the correlations between modalities were captured by learning independent parameters for each modality's specific representation (Zhou et al., 2021b), and these parameters were used to weight and combine modality features to construct correlation representations between modalities. Furthermore a geometric contrastive loss was designed to improve the correlation between missing and complete modalities within the same sample (R. Lin and Hu, 2023). This method functioned by maximizing the similarity between the missing and complete modalities while minimizing the similarity between different samples. However, the reliance on correlation-based methods on assumptions about data distribution may lead to decreased performance when the input data deviates from the expected distribution.

3.1.2.4. Attention-based methods. The attention mechanism originated from studies of human visual perception (X. Li et al., 2023), aiming to simulate how the brain selectively focuses on specific information in complex environments (Y. Zhan & Yang, 2023). In the field of NLP,

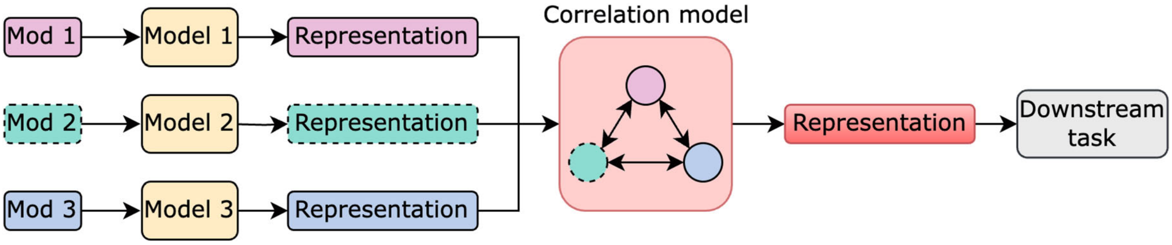


Figure 8. Basic framework of correlation-based methods, with dashed lines indicating the modality is incomplete.

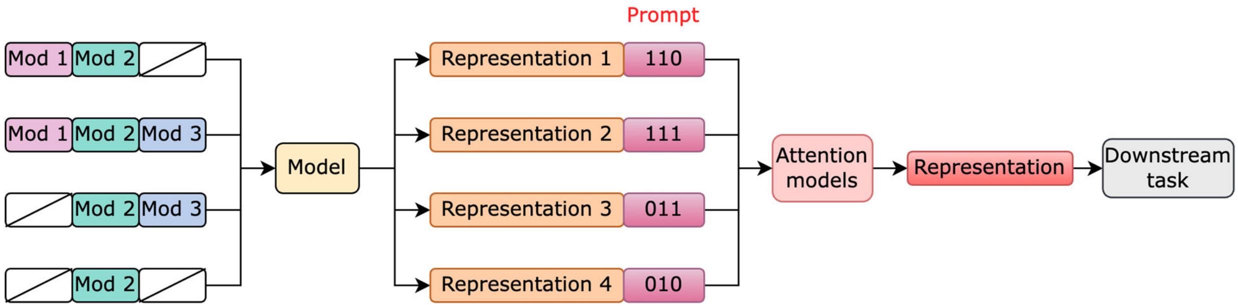


Figure 9. Basic framework of attention-based methods using 0 to indicate missing modality and 1 to indicate available modality.

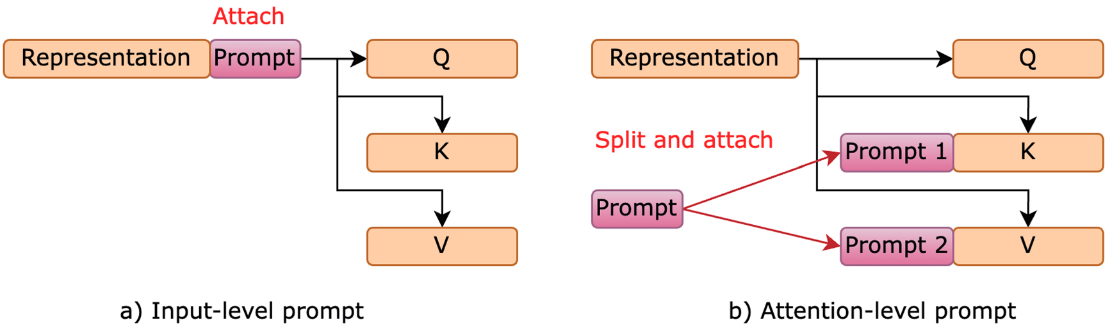


Figure 10. Illustration of two types of prompts.

attention mechanisms are widely used to enhance models' ability to capture key information. They allow models to dynamically assign different weights to various parts of the input, emphasizing important features while ignoring redundant information during data processing (Zeng, Zhou, et al., 2022b). The core idea behind using attention mechanisms to address modality missing issues is to obtain shared latent features by compensating for missing information through a focus on key modalities or features.

In multimodal learning, attention mechanisms automatically assign weights to different modalities, selectively focussing on the most relevant or important ones. Typically, attention-based methods are combined with prompt learning, which guides pre-trained models to perform tasks by designing specific prompts (Min et al., 2023; Qin, Zhang, et al., 2023). In the case of modality missing, prompt learning assigns identifiers (such as 0–1 labels) to indicate which modalities are missing. The model can then use cross-modal attention mechanisms to extract complementary information from other modalities, completing or inferring the missing parts, ensuring the accuracy and robustness of the task, as shown in Figure 9.

Transformer-based models and their attention mechanisms offer significant advantages in addressing modality missing (M. Ma et al., 2022). Since fine-tuning Transformer-based models requires substantial computational resources, it is crucial to develop methods that do not necessitate fine-tuning the pre-trained model. By incorporating prompt learning, Transformer-based models can effectively handle missing modalities by adjusting a small number of learnable parameters, allowing the model to focus on key modalities or features. For instance, a Tag-Assisted Transformer Encoder (TATE) network was proposed in Zeng, Zhou, et al. (2022b) for handling multimodal sentiment analysis with partially missing modalities. This approach introduced a tag encoding module to manage the absence of one or more modalities, guiding the network to focus on the missing modalities and extract the remaining raw features.

The Missing-Aware Prompts were introduced (Lee et al., 2023) to tackle the challenge of missing modalities, requiring adjustments to less than 1% of the learnable parameters (e.g. pooling and fully connected layers). As illustrated in Figure 10, the paper proposed two prompt configuration methods: input-level prompting, which attached prompts to each layer's input, and attention-level prompting, which embedded prompts into the model's attention mechanism. Experimental results indicated that different prompt configurations influenced the effectiveness of learning instructions for pre-trained models. Notably, attention-level prompting was less sensitive to dataset variations and significantly improved baseline performance across various scenarios.

In scenarios where data can be represented as a graph structure, Graph Neural Networks (GNNs) provide a practical framework for addressing the issue of missing modalities (Y. Zhang et al., 2021). GNNs map instances and their features into nodes and edges within a graph, capturing the complex interactions between different modalities. The attention mechanism dynamically adjusts the weights between nodes and their neighbours, allowing the model to focus on key modalities and essential features, thus better compensating for missing information and enhancing overall performance and robustness (Azad et al., 2022; S. Qian & Wang, 2023; Zhou et al., 2023).

Besides, multimodal data was modelled as a Heterogeneous Hypernode Graph (HHG) (J. Chen and Zhang, 2020), with each hypernode representing an instance and hyperedges capturing relationships between instances. The HHG distinguished different modalities and their feature combinations, utilizing Multi-fold Bilevel Graph Attention Networks (MBGAT) to aggregate neighbourhood information from both similar and different modalities. The attention mechanism dynamically assigned weights to nodes based on their importance, allowing the model to focus on key modalities. Furthermore, heterogeneous multimodal data was fused using a dual-stage graph attention network with intra- and inter-modal aggregation mechanisms (Y. Liang, 2024), enhancing interaction between nodes. This approach mapped data to lower-dimensional feature space and learned attention coefficients, allowing nodes to receive key information from neighbouring nodes and modalities.

3.2. External information-based methods

Methods based on external information utilize external resources or knowledge to supplement missing modality information. These approaches can effectively combine information from various sources to address the issue of modality absence (Vapnik & Izmailov, 2015; Wan et al., 2021). Based on the source of external resources, these methods can be categorized into three types: knowledge transfer-based, graph knowledge enhancement-based and human-in-the-loop-based methods.

3.2.1. Knowledge transfer-based methods

These methods focus on transferring knowledge from one domain to another, or from one model to another. Knowledge transfer-based methods include knowledge distillation-based and domain adaptation-based methods.

3.2.1.1. Knowledge distillation-based methods. To address the challenges of applying large models on resource-constrained devices, model compression was introduced in 2006 (Buciluǎ; et al., 2006), allowing simple and fast models to be derived from large, complex, and high-performing models. Inspired by model compression, knowledge distillation was later proposed (Hinton, 2015), which involved training a complex teacher model to obtain accurate results and then transferring its knowledge to a small student model to enhance its performance. In the context of handling missing modalities, knowledge distillation enables the student model to learn from the teacher model's complete multimodal information, compensating for missing modalities to improve its performance on incomplete multimodal datasets, as shown in Figure 11.

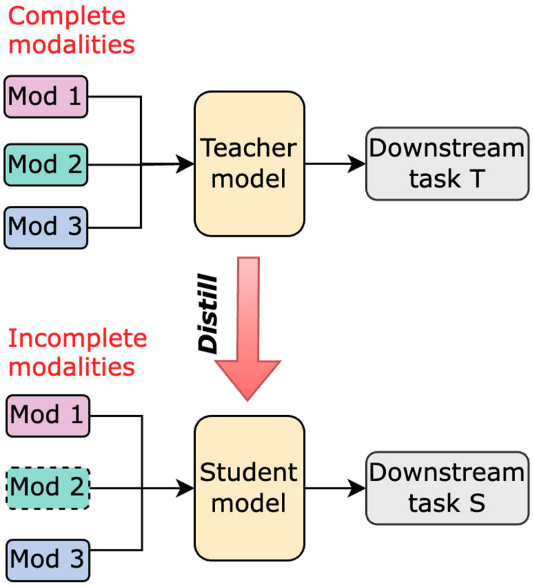


Figure 11. Basic framework of knowledge distillation-based methods, with dashed lines indicating the modality is incomplete.

The KD-Net framework was proposed to improve single-modal performance in medical image segmentation by transferring knowledge from a multimodal teacher network to a single-modal student network (M. Hu et al., 2020). However, its application was constrained as it only accepted single-modal input during testing. In the same field, the Hierarchical Adversarial Knowledge Distillation Network (HAD-Net) was proposed, which reduced the domain gap by mapping segmentation results and features from both networks to a shared space (Vadacchino et al., 2021). However, this method could not generalize to cases of random missing modalities. To handle random missing modalities, the Adversarial Co-training Network (ACN) was proposed (Y. Wang et al., 2021). This network established a coupled learning mechanism that allowed both complete and missing modalities to mutually enhance each other's domain and feature representations. This method enabled the student network to accommodate any arbitrary combination of modality subsets as input during the inference process.

In recent years, Transformers (Qiu et al., 2024) have demonstrated significant advantages in capturing long-range dependencies and global context information. The

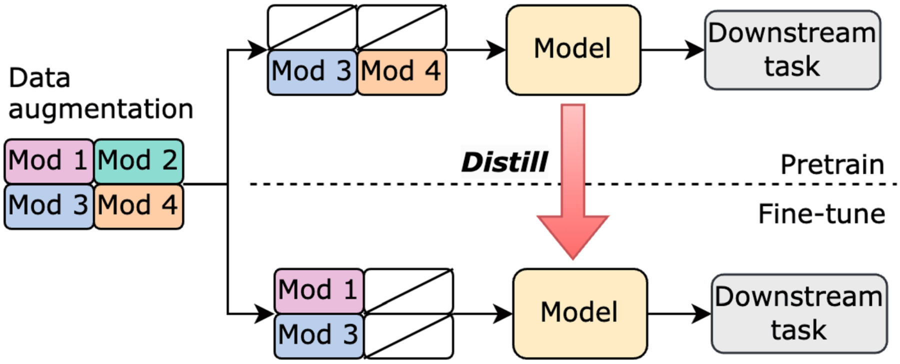


Figure 12. Basic framework of self-distillation in the M^{3}AE.

IMS^{2}Trans network was proposed for brain tumor segmentation (D. Zhang et al., 2024), which employed a shared-weight encoder based on the Swin Transformer, combined with shifted multilayer perceptron and masked bottleneck techniques to effectively capture both local and global information. Through feature distillation, the network was able to accurately extract features even in the case of missing modalities.

The conventional knowledge distillation methods mentioned above rely on pre-trained teacher models to guide the training of student models. If the teacher model performs poorly or overfits, it adversely affects the student model. Additionally, training the teacher model typically requires complete modal data, which may be challenging to obtain in practical applications (Gou et al., 2021). To address these issues, self-distillation leverages the model's own outputs and internal features without requiring a separate teacher model (Ge et al., 2021; M. Ji et al., 2021). This method uses data augmentation to create distinct training datasets, ensuring that the network produces consistent predictions for the same instance or class. Pre-training is conducted on one type of augmented data, while fine-tuning is performed on another. By utilizing its own outputs during training, the model refines its predictions and internal representations, achieving effective self-distillation.

For example, the Multimodal Masked Autoencoder (M^{3}AE) framework was proposed (H. Liu, Wei, et al., 2023), which consisted of pretraining and fine-tuning stages. During pre-training, M^{3}AE used the masked autoencoder principle to reconstruct complete images from partially observed ones, effectively handling missing modalities. By masking 3D blocks to simulate missing data, the framework learned both global and local features. In the fine-tuning stage, a self-distillation strategy enhanced semantic consistency and improved the network's ability to handle missing modalities, as shown in Figure 12.

3.2.1.2. Domain adaptation-based methods. Domain adaptation, a form of transfer learning, addresses the distribution mismatch between training (source domain) and testing (target domain) data in machine learning (X. Chen et al., 2023; Z. Chen, Yang, Huang, Wang et al., 2024). The core idea of domain adaptation-based methods is to leverage existing, complete modality data from the source domain to support learning in the target domain where certain modalities might be missing (X. Chen et al., 2024). Conventional transfer learning methods often focus on transforming modalities within a single dataset or transferring knowledge across datasets from different domains. These methods typically assume that both source and target data are available during training. However, in real-world scenarios, target modality data is often unavailable (X. Chen et al., 2023).

To tackle this issue, a low-rank transfer learning framework called Missing Modality Transfer Learning via Latent Low-Rank Constraint (M^{2}TL) was proposed (Ding et al., 2014), (Ding et al., 2015). This method addressed two challenges of transfer learning: cross-domain transfer, which involved transferring knowledge from one database to another, and cross-modality transfer, which pertained to transferring knowledge from the source modality to the target modality. A basic framework for domain adaptation in this method is shown in Figure 13. The approach maps source and target modality data into a shared subspace using low-rank constraints, enabling data reconstruction and aligning source features with target features. For cross-domain knowledge transfer, the alignment of the two datasets' shared subspaces allowed the effective transfer of knowledge from the source to the target database. Additionally, Maximum Mean Discrepancy (MMD) was used as a regularization term to reduce distribution differences, further improving transfer effectiveness.

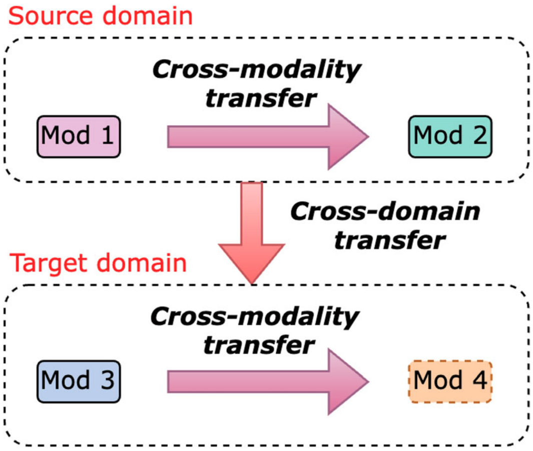


Figure 13. Basic framework of domain adaptation-based methods, with dashed lines indicating the modality is incomplete.

Recently, deep transfer learning has shown clear advantages in addressing missing modalities. Compared to conventional methods, it enhances modality completion and knowledge transfer while capturing complex data patterns, enabling smoother cross-modality transfer (X. Chen et al., 2023). Therefore, a method using random modality dropout training and domain adversarial similarity loss was proposed (Y. Shen and Gao, 2019). Independent encoding paths generated feature representations for each modality, which were fused for segmentation. The model randomly dropped one modality during training, using adversarial loss to maintain consistency between missing and complete modalities, enabling effective brain tumor segmentation despite modality dropouts.

Similarly, the Progressive Modality Cooperation (PMC) method was proposed (W. Zhang et al., 2021), which first trained domain-invariant and modality-specific models using labelled source data and unlabelled target data, then refined target samples with pseudo-labels. A Multi-Modality Generation (MMG) network was introduced to ensure that the generated data maintains domain-invariant characteristics through adversarial learning while preserving semantic information. However, a common issue in domain adaptation-based methods is the distribution difference, as the assumption of correlation between source and target domains may not always hold, and the complexity of missing modalities in the target domain can adversely affect model performance (Wan et al., 2021).

3.2.2. Graph knowledge enhancement-based methods

Knowledge augmentation is a method that enhances model performance during training or inference by incorporating additional external knowledge, such as prior knowledge, knowledge graphs, rules, or domain-specific expertise. Graph-based knowledge augmentation leverages graph structures, such as knowledge graphs, relational graphs, or domain-specific graph structures, to provide external knowledge support for machine learning models. This approach improves the model's reasoning capabilities, enabling it to understand and process complex relationships within the data. These methods build graphs to represent entities, nodes, and the relationships between them, utilizing the information in the graph to enhance feature learning and reasoning. The main approaches include adjacency-based and knowledge graph-based methods (Y. Yang et al., 2022).

3.2.2.1. Adjacency-based methods. The adjacency-based method improves model performance by utilizing the similarity or interaction relationships between different modalities (C. Wang et al., 2016). It fills in missing modalities by retrieving similar or past samples or by propagating features through GNNs, as shown in Figure 14.

Based on task-guided deep kernel functions, a model M^{3}care was proposed (C. Zhang et al., 2022) to compute

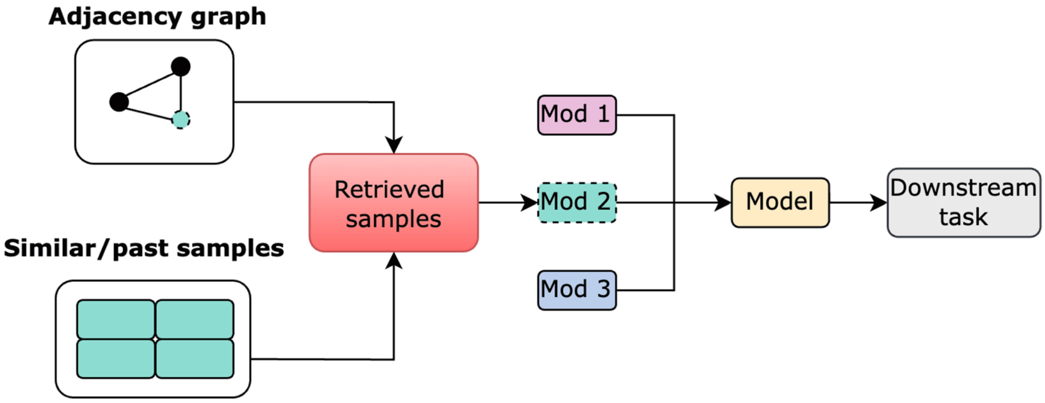


Figure 14. Basic framework of adjacency-based methods, with dashed lines indicating the modality is incomplete.

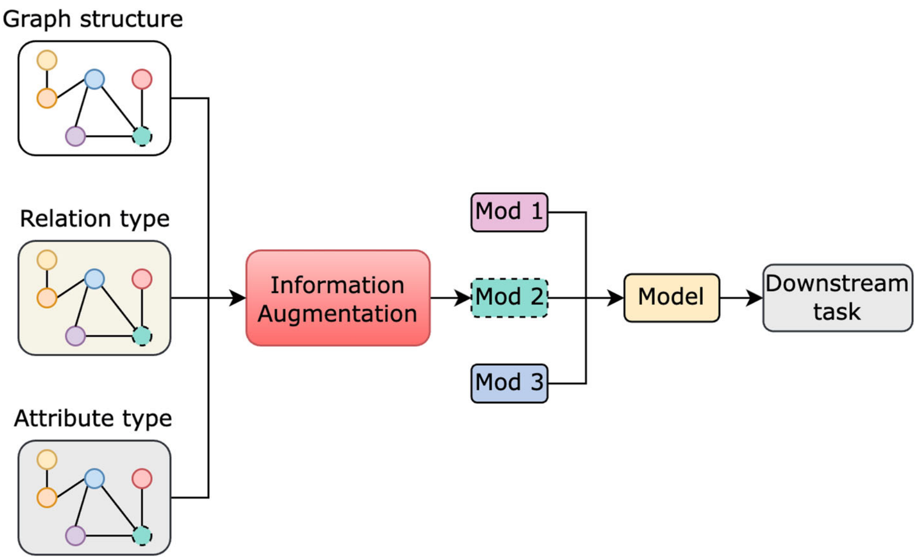


Figure 15. Basic framework of knowledge graph-based methods, with dashed lines indicating the modality is incomplete.

patient similarity, construct a patient graph, and aggregate information from similar patients to adaptively fill missing modalities. Clinical tasks were completed by capturing dynamic intra- and inter-modality interactions. Additionally, the feature propagation method was introduced (Malitesta et al., 2024), which transformed the missing modality problem into a graph node feature missing problem. By iteratively propagating observed features in the graph and optimizing its smoothness, the approach provided complete feature input for subsequent graph neural network learning. Experimental results demonstrated significant effectiveness in multimodal recommendation systems.

3.2.2.2. Knowledge graph-based methods. Knowledge Graphs (KG) represent knowledge using a graph structure, where nodes represent entities and edges represent the relationships between those entities (Hogan et al., 2021). The core idea is to connect different pieces of information through a semantic network, helping machines understand and reason about complex real-world concepts and their interrelationships (C. Peng et al., 2023). The approach to solving the modality missing problem based on KG is to introduce the entities and relationships from the graph to assist in inferring the missing modality features (S. Ji et al., 2021). A basic framework of the KG-based methods is shown in Figure 15.

To reduce the impact of modality loss, the Multi-modal Knowledge Graph Representation Learning (MMKRL), a model that integrated multi-source knowledge through knowledge reconstruction and adversarial training, was proposed (X. Lu et al., 2022). MMKRL embedded structured knowledge and knowledge from modalities such as text and vision into a unified vector space and aligned them using translation methods. At the same time, adversarial training was used to improve the model's ability to handle modality loss and attacks. Experiments showed that MMKRL outperformed other baseline methods in link prediction and triple classification tasks, especially under conditions of limited multi-source knowledge, demonstrating its effectiveness and robustness.

3.2.3. Human-in-the-loop-based methods

Human-in-the-loop (HITL) initially has proven to be a solution to the difficulties of data annotation, where humans select and label key data samples to enhance model performance in specific tasks, particularly in scenarios of data scarcity and high annotation costs (Budd et al., 2021). As technology has advanced, the application scope of HITL has gradually expanded to encompass model training and inference processes (X. Wu et al., 2022). By involving humans in model fine-tuning, error correction, and evaluation, HITL improves the accuracy and robustness of models in complex scenarios. HITL has evolved into a widely applicable framework for data preprocessing, model optimization, and system design (Mosqueira-Rey et al., 2023; Zanzotto, 2019). Additionally, the human feedback mechanism in HITL addresses the shortcomings of deep learning models' 'black box' nature, enhancing models' transparency and interpretability (H. Liu, Yang, et al., 2023). The goal of this human-machine collaboration model is to combine the computational power of machines with human intelligence to achieve higher model accuracy with less data and lower costs (Mosqueira-Rey et al., 2023).

Figure 16 illustrates the iterative method of HITL. When the model's output confidence is low, user intervention becomes particularly important. Human users can review

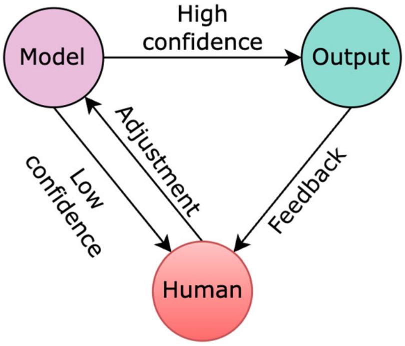


Figure 16. The iterative process of HITL.

the model's outputs based on their own knowledge and experience, providing feedback to correct the model's errors or uncertainties through annotations, comments, and other means (Ouyang et al., 2022). This feedback is incorporated into the model's further training and optimization, achieving iterative improvement. With user feedback, the model can not only enhance specific predictions but also gradually learn the user's preferences and domain knowledge, thereby increasing its own confidence. Continuous optimization of the model is achieved through multiple cycles, allowing the model to gradually adapt and improve its performance in specific tasks (Zanzotto, 2019). Methods for addressing modality missing based on HITL can be classified into two types: one relies on curriculum learning to design tasks, while the other depends on human fine-tuning to optimize the model.

3.2.3.1. Curriculum learning-based methods. In curriculum learning-based methods, human involvement is primarily reflected in the design of the curriculum and the control of training difficulty (Budd et al., 2021). Curriculum learning progressively escalates the complexity of the training data (Bengio et al., 2009; X. Wang et al., 2021), enabling the model to learn to handle missing modalities from simple to complex tasks, as shown in Figure 17.

Human involvement includes designing the curriculum's difficulty curve to help the model gradually master handling missing modalities (Soviany et al., 2022). Experts determine the sequence and complexity of tasks, allowing the model to strengthen its inference abilities progressively. Additionally, human guidance prioritizes which modalities or scenarios are presented during training, preventing the model from facing overly complex tasks too early. Task strategies are also adjusted grounded in the model's feedback (Budd et al., 2021; X. Wang et al., 2021; Zanzotto, 2019). For example, curriculum learning strategies were applied to train Multimodal Generative Adversarial Networks (MM-GANs) (A. Sharma & Hamarneh, 2019). In the experiment, the method began training with simple single-sequence missing scenarios and gradually transitioned to complex ones, resulting in improved model performance. The design of the curriculum learning strategy allowed the model to gradually handle missing modalities as the difficulty of the tasks increased.

3.2.3.2. Human fine-tuning-based methods. Human fine-tuning methods use a feedback loop where users provide real-time input when the model's output is uncertain or missing modalities. Users can correct the model or add missing information based on their knowledge, which helps adjust the model's parameters and improves its ability to infer missing modalities (Qin,

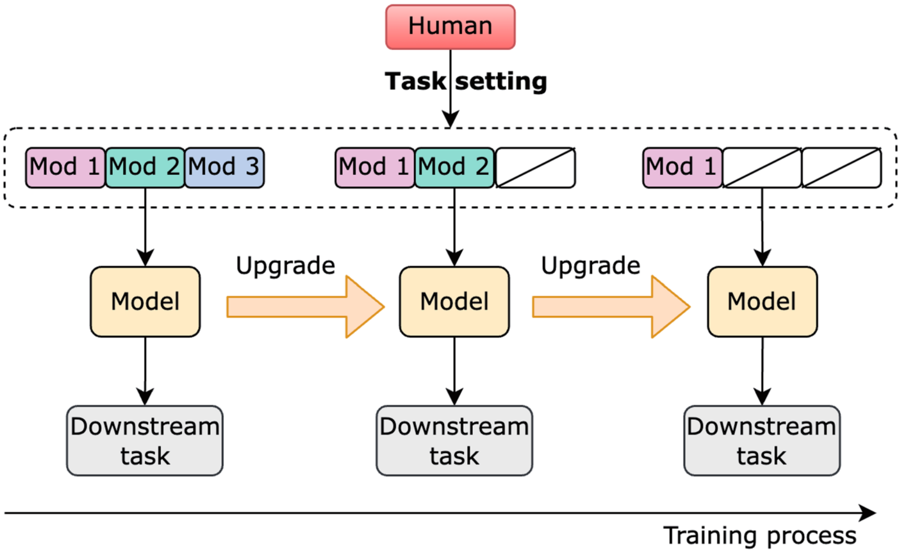


Figure 17. Basic framework of curriculum learning-based methods.

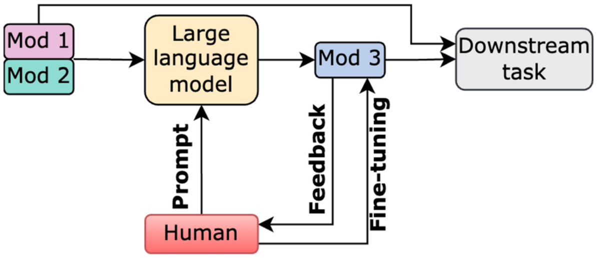


Figure 18. Basic framework of human fine-tuning-based methods.

Zhang, et al., 2023; Zanzotto, 2019). This interactive feedback enhances the model's adaptability and accuracy, promotes informed decision-making, and increases transparency and reliability with multimodal data. Additionally, user feedback can be used for further model training, boosting performance in multimodal learning. A basic framework of human fine-tuning-based methods is shown in Figure 18.

Based on this approach, researchers have proposed the challenge of effectively interpreting model outputs (Qin, Zhang, et al., 2023; Zanzotto, 2019). For instance, the Late-Blind Model used a combination of visual memory and auditory perception in late-blind individuals to help understand situations involving missing modalities (D. Hu et al., 2019). By integrating auditory perception with past visual experiences, the model generated visual content that matches the sounds heard, thereby improving the effectiveness of cross-modal perception. This method demonstrated how human feedback and experience can enhance the model's perceptual abilities in the case of missing modalities.

In remote sensing, data scarcity is a common issue. An innovative approach was proposed that leverages pre-trained large models to generate initial text descriptions for remote sensing images (Ou et al., 2023). By using human feedback to refine these prompts with GPT-4 (Achiam et al., 2023), the model can accurately synthesize remote sensing images. The study shows that by combining a small number of unlabelled images with pre-trained image generation models and refining them through human feedback, the challenges of generating post-disaster remote sensing images can be effectively addressed (R. Mao et al., 2023; Nori et al., 2023). This method not only mitigates the impact of data scarcity but also emphasizes the importance of combining human feedback with large pre-trained models when computational resources are limited.

4. Qualitative analysis

This section analyzes IML-based methods from the perspectives of comparisons among similar methods and comparisons among different methods, with Table 2 summarizing the comparison.

4.1. Comparative analysis within similar methods

4.1.1. Data-based methods

Data synthesis-based methods effectively generate missing modalities by learning the complex distribution and feature representations of the data. This approach overcomes the limitations of missing value imputation techniques, such as zero imputation and mean imputation, which fail to adequately consider the true data distribution (Cai et al., 2018; Pan et al., 2021; Shang et al., 2017; Suo et al., 2019). However, generative models for multimodal data often demand significant computational resources and extended training times. Their generalization capability may also be limited in noisy or heterogeneous datasets, potentially causing the generated data to deviate from the original distribution (Q. Wang, Ding, et al., 2018, 2020; R. Wu et al., 2020; C. Zhang et al., 2020).

Data grouping-based methods reduce noise introduction and information loss by partitioning data into different subsets for independent modelling, which simplifies the challenge of incomplete multimodal alignment (S. Qian & Wang, 2023). However, this approach faces issues such as reduced sample sizes, increased model complexity, and high resource consumption, particularly when scaling in limited or high-dimensional data scenarios. Additionally, there may be sample imbalances among subsets, leading the model to favor groups with larger amounts of data during training. Defining reasonable grouping criteria is also highly challenging for complex multimodal datasets, and it may overlook global feature dependencies between modalities, resulting in insufficient information integration (J. Sun et al., 2024; Xiang et al., 2013; Yuan et al., 2012).

4.1.2. Feature-based methods

First, cyclic translation-based methods can effectively explore the relationships between modalities by maintaining consistency among them (S. Qian & Wang, 2023). This approach promotes the interaction and fusion of features across different modalities. However, ensuring consistency between modalities can pose challenges when dealing with highly heterogeneous data. Furthermore, the results can be influenced by the selection of source and target modalities. Specifically, the choice of source modality can significantly impact the model's performance, as different source modalities provide varying information and features (P. Li et al., 2022; Pham et al., 2019; Q. Wang, Ding, et al., 2020; R. Zhang et al., 2023).

Table 2. Comparison of IML-based methods.

Method Category	Included Subcategories	Included Subcategories	Core Concepts	Advantages	Disadvantages
Internal Information-based	Data-based	Data Synthesis-based	Infer missing modalities using neural network-based generative models.	Complex distribution learning and learning on complete modalities for downstream tasks.	High resource consumption, and limited generalization capability.
		Data Grouping-based	Partition incomplete data by data availability, and train models separately for each group.	Reducing noise introduction, and simplifying incomplete multimodal alignment challenges.	High resource consumption, sample imbalance among subsets, difficulty in defining reasonable grouping criteria, and potential neglect of global feature dependencies between modalities.
	Feature-based	Cyclic Translation-based	Ensure cyclical consistency between modalities to obtain a shared multimodal latent representation.	Promotion of interaction and fusion between different modality features.	Variation in results due to different source and target modality selections.
		Feature Disentanglement-based	Construct a shared multimodal latent representation by extracting modality-invariant and modality-specific features.	Reduction of interference between different modalities.	High computational costs and training difficulties with high noise and heterogeneous data.
		Correlation-based	Model relationships between samples or modalities to achieve a shared multimodal latent representation.	Task-specific model design with enhanced applicability.	High computational complexity and sensitivity to data distribution assumptions.
		Attention-based	Focus on key modalities or features to compensate for missing information and obtain a shared multimodal latent representation.	Dynamically weighting and focussing on key modalities, and flexibly handling missing modalities.	Susceptibility to the quality of pre-trained models.
External Information-based	Knowledge Transfer-based	Knowledge Distillation-based	Enable the student model to acquire from the complete multimodal information of the teacher model.	Reduction of student model complexity and self-distillation without reliance on an independent teacher model.	Susceptibility to the quality of teacher model.
		Domain Adaptation-based	Use complete modality data (source domain) to handle missing modality data (target domain).	Reduction of labelling requirements.	Susceptibility to data distribution differences.
	Graph Knowledge Enhancement-based	Adjacency-based	Leverage similarity or interaction relationships.	Easy to implement.	Not suitable for complex data.
		Knowledge Graph-based	Connect information via a semantic network to help machines understand complex concepts and relationships.	Enhanced context understanding.	Complex construction and low inference efficiency.
	Human-in-the-loop-based	Curriculum Learning-based	Design the curriculum and control training difficulty, gradually increasing task complexity.	Enhancing model generalization ability.	Susceptibility to the curriculum design.
		Human Fine-tuning-based	Allow users to supplement information or adjust model output.	Allowing real-time learning and adjustment of the model during practical use.	Susceptibility of model performance to feedback quality.

Second, feature disentanglement-based methods effectively extract both modality-invariant and modality-specific features. By employing independent disentanglement, these methods can reduce interference between different modalities, enhancing the model's capacity to learn features unique to each modality. This approach excels in facilitating the transfer of learned features to new tasks, allowing for efficient adaptation and good generalization. However, the processes of feature extraction and fusion can be complex, potentially increasing computational costs and training difficulties, especially when dealing with high noise or heterogeneous data (Y. Chen et al., 2023; R. Liu et al., 2024; H. Wang et al., 2023; Q. Yang et al., 2022).

Third, correlation-based methods emphasize the interrelationships among modalities, capturing potential complex relationships. These methods can design relevant models according to the downstream task, thereby enhancing the effectiveness and applicability of the task. However, they tend to have high computational complexity, particularly when processing large-scale data, which may limit their real-time application. Furthermore, their sensitivity to data distribution assumptions can cause performance degradation when the input does not align with the expected distribution (Y. Liu, Fan, et al., 2021; F. Ma et al., 2021; Q. Wang, Lian, et al., 2020; Zhou et al., 2021b).

Finally, attention-based methods dynamically assign weights to effectively focus on key modalities. This approach combines with prompt learning to flexibly address missing modalities. However, its performance is reliant on the quality of the pre-trained model. If the pre-trained model performs poorly on specific modalities, it may result in insufficient information extraction, and the choice of prompts can also affect the model's adaptability to missing modalities (J. Chen & Zhang, 2020; X. Li et al., 2023; Y. Zhan & Yang, 2023).

4.1.3. Knowledge transfer-based methods

Knowledge distillation-based methods effectively help student models enhance performance on incomplete multimodal datasets and improve their adaptability on resource-constrained devices by extracting knowledge from complex teacher models (Q. Wang, Zhan, et al., 2020). However, the performance of the student model is highly dependent on the quality of the teacher model. It is assumed that the teacher model possesses complete modality information, which may be difficult to achieve in practical applications (Vadacchino et al., 2021). Self-distillation does not rely on an independent teacher model (Ge et al., 2021; M. Ji et al., 2021), but the training process may still face high computational costs, especially when dealing with large-scale datasets.

Domain adaptation-based methods leverage relevant source domain data to support target domain learning, providing new possibilities for addressing missing modalities in the target domain (Kim et al., 2022). This approach enhances model adaptability and lowers annotation requirements. However, the data from the source and target domains may not necessarily follow the same distribution, making it difficult to generalize to the complex scenarios of modality missing in the target domain (Ding et al., 2015).

4.1.4. Graph knowledge enhancement-based methods

Adjacency-based methods are relatively easy to implement and exhibit adaptability in cases of sparse data (C. Wang et al., 2016). However, they face challenges in capturing complex relationships within the data, particularly when interactions among modalities are intricate or when latent factors influence the observed features (Malitesta et al., 2024; Y. Yang et al., 2022; C. Zhang et al., 2022).

Knowledge graph-based methods utilize semantic information to provide a structured representation of information, clarifying relationships between modalities (Budd et al., 2021). However, constructing high-quality knowledge graphs requires extensive manual annotation and domain expertise, leading to high costs and potential delays in updates. Additionally, knowledge graph inference can be computationally intensive, particularly with large-scale graphs. If the knowledge graph lacks relevant entities or relationships, inference outcomes may also be limited (Hogan et al., 2021; X. Lu et al., 2022; C. Peng et al., 2023; X. Wu et al., 2022).

4.1.5. Human-in-the-loop-based methods

Curriculum learning-based methods effectively guide the model's learning process, reducing cognitive load during training and enabling the model to perform robustly on complex tasks. However, designing an appropriate curriculum requires substantial domain knowledge and experience, and a poorly designed curriculum can result in suboptimal learning outcomes. Furthermore, an over-reliance on curriculum design may limit the model's adaptability in handling unseen, complex situations, potentially constraining its generalization capacity.

Human fine-tuning-based methods use a feedback loop that allows users to provide real-time input when model outputs are uncertain or modalities are missing. This enables real-time learning and adjustment of the model during practical use. Nevertheless, the method's effectiveness is reliant on the quality of user feedback, and inaccuracies can lead to erroneous learning. Additionally, real-time feedback may burden users, especially with large datasets.

4.2. Comparative analysis between different methods

4.2.1. Comparison of internal information-based and external information-based methods

The main advantage of internal information-based methods lies in efficiently utilizing internal data without needing additional external data collection or integration. This reliance on internal data also enhances data privacy, thereby lowering privacy risks and ensuring compliance with data protection regulations (J. Chen & Zhang, 2020; Lee et al., 2023; Y. Zhan & Yang, 2023). Additionally, they reduce the risk of data shift that arises from differing distributions or biases in external sources, leading to stable model performance in inference. However, their limitation is that if the dataset lacks sufficient information or has uneven sample distributions, the model's inference capabilities may be restricted (Y. Liu, Fan, et al., 2021; F. Ma et al., 2021; Q. Wang, Lian, et al., 2020; Zhou et al., 2021b).

In contrast, external information methods leverage outside resources, such as contextual information, extra datasets, domain knowledge, or human feedback. Accessing diverse information sources improves inter-modal understanding and prediction accuracy, while human feedback allows dynamic adjustments of models, boosting adaptability (Kim et al., 2022). However, these methods face challenges in acquiring and integrating external resources, adding complexity and time costs. Low-quality information may introduce noise, and reliance on human input can limit scalability, especially in large-scale applications (Malitesta et al., 2024; Q. Wang, Zhan, et al., 2020; Y. Yang et al., 2022; C. Zhang et al., 2022).

4.2.2. Comparison of data-based and feature-based methods

Data-based and feature-based methods pay attention to different aspects when addressing the issue of missing modalities. Data-based methods emphasize modality completeness, typically by generating missing modalities or grouping data to ensure downstream tasks operate on fully complete modalities (Q. Wang, Ding, et al., 2018, 2020). In contrast, feature-based methods place greater emphasis on integrating existing modality information. Through feature extraction, integration, and optimization, they leverage available modality features for reasoning and decision-making to compensate for missing modalities (W. Sun et al., 2021; Q. Wang, Ding, et al., 2020).

4.2.3. Comparison of knowledge transfer-based and graph knowledge enhancement-based methods

Knowledge transfer-based methods rely directly on existing models or data sources (such as teacher models or source domain data) to transfer knowledge. These methods are effective for handling similar tasks, but their performance is heavily dependent on the quality of the teacher model and the similarity between the two domains (Ge et al., 2021). If the teacher or source domain model is of poor quality, the effectiveness of the transfer can be significantly compromised (Ding et al., 2015). In contrast, graph knowledge enhancement-based methods utilize mechanisms such as knowledge reasoning and relational inference, emphasizing reasoning and integration from external knowledge bases (such as knowledge graphs) and structured information (Hogan et al., 2021). These methods focus on multi-dimensional information integration, enabling them to handle diverse and complex data relationships (Hogan et al., 2021; X. Lu et al., 2022; C. Peng et al., 2023; X. Wu et al., 2022).

4.2.4. Comparison of curriculum learning-based and data grouping-based methods

Curriculum learning-based methods emphasize continuous learning and dynamic feedback during training. With a human-designed curriculum, the model receives feedback on simple tasks and gradually advances to complex ones. This approach allows real-time adjustments to the learning strategy, enhancing adaptability to modality gaps. It focuses on the growth of a single model within a dynamic environment, characterized by ongoing optimization throughout the learning process (Graves et al., 2017; Pentina et al., 2015). In contrast, data grouping-based methods prioritize independent training across multiple models or data subsets instead of continuous learning in a single model. Training relies on the distributional characteristics of existing data, without direct human influence on the learning trajectory. This method emphasizes collaboration among models, making it suitable for handling static multimodal datasets. In cases of significant feature variation, data grouping leverages independent modelling to enhance robustness (J. Sun et al., 2024; Xiang et al., 2013; Yuan et al., 2012).

5. Challenges and opportunities

The issue of modality missing has long been a critical topic in multimodal learning. Early studies primarily focussed on methods such as data imputation and data generation. Recently, with the widespread utilization of neural networks, research in IML has made significant progress with an increasing number of studies aiming to tackle modality missing from various perspectives. However, in real-world open environments, the transition of IML techniques from research to widespread application still faces many challenges. The following sections discuss the technical difficulties and future research directions in this field.

5.1. Modality heterogeneity

In modality missing research, modality heterogeneity poses a significant challenge, primarily due to variability and inconsistency across different data sources or features. First, each modality often follows its unique statistical distribution – for instance, image data tends to exhibit high nonlinearity and high-dimensional features, while text data involves linguistic diversity and complex grammar structures (Zheng et al., 2023). These distributional differences can lead to model bias, affecting prediction accuracy. Additionally, differences in measurement scales, such as between numerical and categorical data, can impact the effectiveness of conventional data imputation methods (J. Chen & Zhang, 2020). Feature importance also varies across modalities, requiring careful feature selection and weighting when handling missing data. Complex inter-modal associations further complicate processing, especially when one modality is missing and others cannot fully compensate (W. Liang et al., 2021).

5.2. Flexibly addressing random and imbalance missingness in both training and testing phase

The ultimate goal of IML is to effectively learn from incomplete multimodal training sets while ensuring that the model can dynamically and robustly handle new instances with any missing modalities during the testing phase. This means that the model needs to be flexibly designed within a unified framework to handle various modality-missing scenarios during both phases, ensuring good performance under different modality conditions. To overcome these limitations, future models for IML should be designed to dynamically adjust the modalities utilized during both phases. This involves implementing adaptive strategies that enable the model to evaluate the availability and relevance of each modality in real time (Z. Liu et al., 2024; Y. Sun et al., 2024; Zeng, Zhou, et al., 2022b).

5.3. Cross-domain modality alignment and mapping

Cross-domain modality alignment and mapping in IML encounter significant challenges, especially when data may be scarce. Models need to effectively complete and generalize missing modalities, which require techniques like self-supervised learning and few-shot learning to extract useful cross-domain features (Xue et al., 2024; Y. Zhang et al., 2022). Additionally, designing robust alignment mechanisms is crucial to avoid overfitting while accurately mapping modality relationships across different domains. Conventional domain adaptation-based methods struggle in scenarios where certain modalities are entirely missing (C. Yang et al., 2022). Therefore, developing domain-invariant feature alignment strategies is essential to enable models to learn shared features in the presence of modality missing.

5.4. Catastrophic forgetting in curriculum learning

Addressing the missing modality requires immediate compensation strategies. It is also important to consider catastrophic forgetting, which refers to the rapid forgetting of previously acquired information when new data is introduced (Goodfellow et al., 2013). In dynamic environments, the periodic absence of modalities can disrupt the learning process, making long-term learning and memory mechanisms crucial (Y. Zhan et al., 2024). These models are required to learn missing patterns from historical data and predict future absences, necessitating algorithms with temporal dependencies and memory capabilities. In deep learning frameworks, exploring new long-term memory mechanisms is essential to mitigate knowledge forgetting (L. Wang, X. Zhang, et al., 2024; Zhu et al., 2023).

5.5. Accurate understanding of user feedback

In the interaction between users and large language models, accurately inferring the emotions and tone in user feedback, as well as avoiding misunderstandings, poses a significant challenge in guiding the model to generate missing modalities (Cao et al., 2023; Mei et al., 2011; Wei et al., 2023). User feedback is influenced by cultural backgrounds, personal experiences, and subjective emotions, which can lead to vague or inconsistent emotional signals. Therefore, the model needs to be sensitive and flexible in capturing emotional changes and tonal nuances (Axelsson & Skantze, 2022). Thus, users not only act as feedback providers but also need to actively guide the model by clearly expressing emotional intentions and providing contextual information to enhance communication effectiveness and improve the model's responsiveness.

5.6. Dynamic personalization models in federated learning

In federated learning, clients frequently encounter varying degrees of modality missing, which necessitates personalized adjustments tailored to each client's data distribution. Dynamic personalized models adjust model parameters in response to missing modalities to maintain adaptability and accuracy in changing environments (Zheng et al., 2023). However, implementing dynamic personalized models requires real-time updates to address modality changes and needs privacy-preserving monitoring mechanisms, which place high demands on communication efficiency. Future research should focus on developing dynamic feedback mechanisms, adaptive optimization algorithms, and cross-modal learning frameworks to personalize the approach to tackle modality missing (S. Yu et al., 2024).

5.7. Application to a wide range of intelligent tasks

Existing research on IML has primarily focussed on tasks such as medical image segmentation and generation, as well as sentiment analysis (X. Chen et al., 2024; Z. Lu & Guo, 2023). However, in practical applications, other intelligent tasks – such as fault diagnosis, autonomous driving, environmental monitoring, and industrial inspection – also require the capability to handle missing modality issues (S. Ma & Li, 2023; Z. Wu et al., 2024; Zhong et al., 2023). When designing IML algorithms for these tasks, it is essential to not only draw on existing approaches but also to incorporate domain-specific knowledge to develop models and algorithms with strong relevance (Xue et al., 2024).

6. Conclusion

The challenge of IML is how to effectively learn from incomplete multimodal training sets and ensure that the model can dynamically and robustly handle new instances with arbitrary missing modalities during the testing phase. This paper first reviews the development of multimodal learning, clarifies the challenge of IML, and defines the issue of modality missing. Next, it provides a review of the latest advancements in IML from various technical perspectives, assessing the advantages and disadvantages of different methods. A qualitative analysis of existing methods is conducted from two angles: comparative analysis within similar methods and comparative analysis between different methods, along with suggestions for future research directions. Currently, preliminary explorations in IML have emerged, with further research and development potential in fields such as federated learning, class incremental learning, and transfer learning. Moreover, future IML should target more downstream tasks, including fault diagnosis, autonomous driving, environmental monitoring, and industrial inspection.

Author contributions statement

The following outlines the contributions of all authors. All authors have read and approved the final version of the manuscript:

Yifan Zhan: Conceptualization, Data Curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing–Original Draft, Writing–Review & Editing.

Rui Yang: Conceptualization, Formal analysis, Funding acquisition, Project Administration, Resources, Supervision, Writing–Original Draft, Writing–Review & Editing.

Junxian You: Validation, Visualization, Writing–Original Draft, Writing–Review & Editing.

Mengjie Huang: Conceptualization, Funding acquisition, Resources, Supervision, Validation, Writing-Review & Editing.

Weibo Liu: Supervision, Writing–Original Draft, Writing–Review & Editing.

Xiaohui Liu: Supervision, Writing–Original Draft, Writing–Review & Editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by National Natural Science Foundation of China (72401233), Jiangsu Provincial Qinglan Project, Natural Science Foundation of Jiangsu Higher Education Institutions of China (23KJB520038), Suzhou Science and Technology Programme (SYG202106), and Research Enhancement Fund of XJTLU (REF-23-01-008).

Data availability

The data that support the findings of this study are available from the corresponding author, R.Y., upon reasonable request.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., & Avila, R. (2023). GPT-4 technical report. *Preprint*. arXiv:2303.08774.

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic speech recognition: Systematic literature review. *IEEE Access*, *9*, 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

Axelsson, A., & Skantze, G. (2022). Multimodal user feedback during adaptive robot-human presentations. *Frontiers in Computer Science*, *3*, 741148. https://doi.org/10.3389/fcomp.2021.741148

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E. K., Molaei, A., Jia, Y., Jose, A., Roy, R., & Merhof, D. (2024). Advances in medical image analysis with vision transformers: A comprehensive review. *Medical Image Analysis*, *91*, 103000. https://doi.org/10.1016/j.media.2023.103000

Azad, R., Khosravi, N., Dehghanmanshadi, M., Cohen-Adad, J., & Merhof, D. (2022). Medical image segmentation on

MRI images with missing modalities: A review. *Preprint*. arXiv:2203.06217.

Bayoudh, K., Knani, R., Hamdaoui, F., & Mtibaa, A. (2022). A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *The Visual Computer*, *38*(8), 2939–2970. https://doi.org/10.1007/s00371-021-02166-7

Ben-Cohen, A., Klang, E., Raskin, S. P., Soffer, S., Ben-Haim, S., Konen, E., Amitai, M. M., & Greenspan, H. (2019). Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection. *Engineering Applications of Artificial Intelligence*, *78*, 186–194. https://doi.org/10.1016/j.engappai.2018.11.013

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 41–48).

Buciluǎ, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 535–541).

Budd, S., Robinson, E. C., & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, *71*, 102062. https://doi.org/10.1016/j.media.2021.102062

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).

Cao, Q., Yu, H., Charisse, P., Qiao, S., & Stevens, B. (2023 Mar.). Is high-fidelity important for human-like virtual avatars in human computer interactions? *International Journal of Network Dynamics and Intelligence*, *2*(1), 15–23. https://doi.org/10.53941/ijndi0201008

Chen, C., Dou, Q., Jin, Y., Liu, Q., & Heng, P. A. (2021). Learning with privileged multimodal knowledge for unimodal segmentation. *IEEE Transactions on Medical Imaging*, *41*(3), 621–632. https://doi.org/10.1109/TMI.2021.3119385

Chen, Z., Guo, L., Fang, Y., Zhang, Y., Chen, J., Pan, J. Z., Li, Y., Chen, H., & Zhang, W. (2023). Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In *International Semantic Web Conference* (pp. 121–139).

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., & Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, *13*(22), 4712. https://doi.org/10.3390/rs13224712

Chen, Y., Pan, Y., Xia, Y., & Yuan, Y. (2023). Disentangle first, then distill: A unified framework for missing modality imputation and Alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging*, *42*(12), 3566–3578.

Chen, Z., Yang, R., Huang, M., Li, F., Lu, G., & Wang, Z. (2024). EEGProgress: A fast and lightweight progressive convolution architecture for EEG classification. *Computers in Biology and Medicine*, *169*, 107901. https://doi.org/10.1016/j.compbiomed.2023.107901

Chen, Z., Yang, R., Huang, M., Wang, Z., & Liu, X. (2024). Electrode domain adaptation network: Minimizing the difference across electrodes in single-source to single-target motor imagery classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *8*(2), 1–15.

Chen, X., Yang, R., Xue, Y., Huang, M., Ferrero, R., & Wang, Z. (2023). Deep transfer learning for bearing fault diagnosis: A systematic review since 2016. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–21.

Chen, X., Yang, R., Xue, Y., Song, B., & Wang, Z. (2024). TFPred: Learning discriminative representations from unlabeled data for few-label rotating machinery fault diagnosis. *Control Engineering Practice*, *146*, 105900. https://doi.org/10.1016/j.conengprac.2024.105900

Chen, J., & Zhang, A. (2020). HGMF: Heterogeneous graph-based fusion for multimodal data with incompleteness. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1295–1305).

Croitoru, F.-A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10850–10869. https://doi.org/10.1109/TPAMI.2023.3261988

Dalmaz, O., Yurt, M., & Çukur, T. (2022). ResViT: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, *41*(10), 2598–2614. https://doi.org/10.1109/TMI.2022.3167808

Ding, Z., Ming, S., & Fu, Y. (2014). Latent low-rank transfer subspace learning for missing modality recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 28(1)).

Ding, Z., Shao, M., & Fu, Y. (2015). Missing modality transfer learning via latent low-rank constraint. *IEEE Transactions on Image Processing*, *24*(11), 4322–4334. https://doi.org/10.1109/TIP.2015.2462023

Diwan, T., Anirudh, G., & Tembhurne, J. V. (2023). Object detection using YOLO: Challenges, architectural successors, datasets and applications. *Multimedia Tools and Applications*, *82*(6), 9243–9275. https://doi.org/10.1007/s11042-022-13644-y

Dong, A., Starr, A., & Zhao, Y. (2023). Neural network-based parametric system identification: A review. *International Journal of Systems Science*, *54*(13), 2676–2688. https://doi.org/10.1080/00207721.2023.2241957

Ge, Y., Zhang, X., Choi, C. L., Cheung, K. C., Zhao, P., Zhu, F., Wang, X., Zhao, R., & Li, H. (2021). Self-distillation with batch knowledge ensembling improves imagenet classification. *Preprint*. arXiv:2104.13298.

Ghandi, T., Pourreza, H., & Mahyar, H. (2023). Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, *56*(3), 1–39. https://doi.org/10.1145/3617592

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., & Bengio, Y. (2013). An empirical investigation of catastrophic forgetting in gradient-based neural networks. *Preprint*. arXiv:1312.6211.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, *27*, 2672–2680.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, *129*(6), 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

Graves, A., Bellemare, M. G., Menick, J., Munos, R., & Kavukcuoglu, K. (2017). Automated curriculum learning for neural networks. In *International Conference on Machine Learning* (pp. 1311–1320).

Han, F., Liu, J., Li, J., Song, J., Wang, M., & Zhang, Y. (2023). Consensus control for multi-rate multi-agent systems with fading measurements: The dynamic event-triggered case. *Systems Science & Control Engineering*, *11*(1), 2158959. https://doi.org/10.1080/21642583.2022.2158959

Han, J., Zhang, Z., Ren, Z., & Schuller, B. (2019). Implicit fusion by joint audiovisual training for emotion recognition in mono modality. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5861–5865).

Hao, H., Xue, J., Huang, P., Ren, L., & Li, D. (2024). QGFormer: Queries-guided transformer for flexible medical image synthesis with domain missing. *Expert Systems with Applications*, *247*, 123318. https://doi.org/10.1016/j.eswa.2024.123318

Hinton, G. (2015). Distilling the knowledge in a neural network. *Preprint* arXiv:1503.02531.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., & Ngomo, A. C. N. (2021). Knowledge graphs. *ACM Computing Surveys*, *54*(4), 1–37. https://doi.org/10.1145/3447772

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, *51*(6), 1–36. https://doi.org/10.1145/3295748

Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., & Gori, P. (2020). Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference* (pp. 772–781).

Hu, D., Wang, D., Li, X., Nie, F., & Wang, Q. (2019). Listen to the image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7972–7981).

Huang, Y., Shao, L., & Frangi, A. F. (2017). Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning. *IEEE Transactions on Medical Imaging*, *37*(3), 815–827. https://doi.org/10.1109/TMI.2017.2781192

Islam, M., Wijethilake, N., & Ren, H. (2021). Glioblastoma multiforme prognosis: MRI missing modality generation, segmentation and radiogenomic survival prediction. *Computerized Medical Imaging and Graphics*, *91*, 101906. https://doi.org/10.1016/j.compmedimag.2021.101906

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Philip, S. Y. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

Ji, M., Shin, S., Hwang, S., Park, G., & Moon, I.-C. (2021). Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10664–10673).

Joseph, K., Khan, S., Khan, F. S., & Balasubramanian, V. N. (2021). Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5830–5840).

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139–172. https://doi.org/10.1080/23270012.2020.1756939

Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. *BMC Medical Imaging*, *22*(1), 69. https://doi.org/10.1186/s12880-022-00793-7

Kingma, D. P. (2013). Auto-encoding variational bayes. *Preprint*. arXiv:1312.6114.

Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, *470*, 443–456. https://doi.org/10.1016/j.neucom.2021.05.103

Lee, Y.-L., Tsai, Y.-H., Chiu, W.-C., & Lee, C.-Y. (2023). Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14943–14952).

Li, M., Huang, S.-L., & Zhang, L. (2022). A general framework for incomplete cross-modal retrieval with missing labels and missing modalities. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4763–4767).

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. https://doi.org/10.1109/TII.2022.3197201

Li, B., Li, C., Duan, F., Zheng, N., & Zhao, Q. (2020). TPFN: Applying outer product along time to multimodal sentiment analysis fusion on incomplete data. In *Computer Vision–ECCV 2020: 16th European Conference* (pp. 431–447).

Li, J., Li, L., Sun, R., Yuan, G., Wang, S., & Sun, S. (2024). MMAN-M2: Multiple multi-head attentions network based on encoder with missing modalities. *Pattern Recognition Letters*, *177*, 110–120. https://doi.org/10.1016/j.patrec.2023.11.029

Li, X., Li, M., Yan, P., Li, G., Jiang, Y., Luo, H., & Yin, S. (2023). Deep learning attention mechanism in medical image analysis: Basics and beyonds. *International Journal of Network Dynamics and Intelligence*, *2*(1), 93–116. https://doi.org/10.53941/ijndi0201006.

Li, G., Wang, W., Zhang, W., Wang, Z., Tu, H., & You, W. (2021). Grid search based multi-population particle swarm optimization algorithm for multimodal multi-objective optimization. *Swarm and Evolutionary Computation*, *62*, 100843. https://doi.org/10.1016/j.swevo.2021.100843

Liang, Y. (2024). Multimodal knowledge graph embedding with missing data integration. *IEEE Transactions on Computational Social Systems*, 1–13.

Liang, W., Wang, G., Lai, J., & Xie, X. (2021). Homogeneous-to-heterogeneous: Unsupervised learning for RGB-infrared person re-identification. *IEEE Transactions on Image Processing*, *30*, 6392–6407. https://doi.org/10.1109/TIP.2021.3092578

Liang, P. P., Zadeh, A., & Morency, L.-P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, *56*(10), 1–42.

Lin, R., & Hu, H. (2023). MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, *11*, 1686–1702. https://doi.org/10.1162/tacl_a_00628

Lin, C.-T., Wu, Y.-Y., Hsu, P.-H., & Lai, S.-H. (2020). Multimodal structure-consistent image-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34(7), pp. 11490–11498).

Liu, Y., Fan, L., Zhang, C., Zhou, T., Xiao, Z., Geng, L., & Shen, D. (2021). Incomplete multi-modal representation learning for Alzheimer's disease diagnosis. *Medical Image Analysis*, *69*, 101953. https://doi.org/10.1016/j.media.2020.101953

Liu, J., Pasumarthi, S., Duffy, B., Gong, E., Datta, K., & Zaharchuk, G. (2023). One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. *IEEE Transactions on Medical Imaging*, *42*(9), 2577–2591. https://doi.org/10.1109/TMI.2023.3261707

Liu, Y., Sun, P., Wergeles, N., & Shang, Y. (2021). A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications*, *172*, 114602. https://doi.org/10.1016/j.eswa.2021.114602

Liu, H., Wei, D., Lu, D., Sun, J., Wang, L., & Zheng, Y. (2023). M3AE: Multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37(2), pp. 1657–1665).

Liu, H., Yang, X., & Xu, C. (2023). Counterfactual scenario-relevant knowledge-enriched multi-modal emotion reasoning. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(5s), 1–25. https://doi.org/10.1145/3583690

Liu, Z., Zhou, B., Chu, D., Sun, Y., & Meng, L. (2024). Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, *101*, 101973. https://doi.org/10.1016/j.inffus.2023.101973

Liu, R., Zuo, H., Lian, Z., Schuller, B. W., & Li, H. (2024). Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*, *15*(4), 1856–1873.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3431–3440).

Lu, Z., & Guo, G. (2023). Control and communication scheduling co-design for networked control systems: A survey. *International Journal of Systems Science*, *54*(1), 189–203. https://doi.org/10.1080/00207721.2022.2097332

Lu, X., Wang, L., Jiang, Z., He, S., & Liu, S. (2022). MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Applied Intelligence*, *52*(7), 7480–7497.

Ma, S., & Li, Y. (2023). Adaptive fuzzy fault-tolerant control for active seat suspension systems with full-state constraints. *Systems Science & Control Engineering*, *11*(1), 2153391. https://doi.org/10.1080/21642583.2022.2153391

Ma, M., Ren, J., Zhao, L., Testuggine, D., & Peng, X. (2022). Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18177–18186).

Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., & Peng, X. (2021). Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35(3), pp. 2302–2310).

Ma, F., Xu, X., Huang, S.-L., & Zhang, L. (2021). Maximum likelihood estimation for multimodal learning with missing modality. *Preprint*. arXiv:2108.10513.

Maheshwari, H., Liu, Y.-C., & Kira, Z. (2024). Missing modality robustness in semi-supervised multi-modal semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1020–1030).

Malitesta, D., Rossi, E., Pomo, C., Malliaros, F. D., & Di Noia, T. (2024). Dealing with missing modalities in multimodal recommendation: A feature propagation-based approach. *Preprint*. arXiv:2403.19841.

Mao, R., Chen, G., Zhang, X., Guerin, F., & Cambria, E. (2023). GPTEval: A survey on assessments of ChatGPT and GPT-4. *Preprint*. arXiv:2308.12488.

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*.

Mei, T., Yang, B., Hua, X.-S., & Li, S. (2011). Contextual video recommendation by multimodal relevance and user feedback. *ACM Transactions on Information Systems*, *29*(2), 1–24. https://doi.org/10.1145/1961209.1961213

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, *56*(2), 1–40. https://doi.org/10.1145/3605943

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W.. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551.

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. *Preprint*. arXiv:2303.13375.

Oruh, J., Viriri, S., & Adegun, A. (2022). Long short-term memory recurrent neural network for automatic speech recognition. *IEEE Access*, *10*, 30069–30079. https://doi.org/10.1109/ACCESS.2022.3159339

Ou, R., Yan, H., Wu, M., & Zhang, C. (2023). A method of efficient synthesizing post-disaster remote sensing image with diffusion model and LLM. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference* (pp. 1549–1555).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., & Schulman, J. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pan, Y., Liu, M., Xia, Y., & Shen, D. (2021). Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6839–6853. https://doi.org/10.1109/TPAMI.2021.3091214

Peng, C., Xia, F., Naseriparsa, M., & Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, *56*(11), 13071–13102. https://doi.org/10.1007/s10462-023-10465-9

Pentina, A., Sharmanska, V., & Lampert, C. H. (2015). Curriculum learning of multiple tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5492–5500).

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. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33(1), pp. 6892–6899).

Qian, X., & Cui, B. (2023). A mobile sensing approach to distributed consensus filtering of 2D stochastic nonlinear parabolic systems with disturbances. *Systems Science & Control Engineering*, *11*(1), 2167885. https://doi.org/10.1080/21642583.2023.2167885

Qian, S., & Wang, C. (2023). COM: Contrastive masked-attention model for incomplete multimodal learning. *Neural Networks*, *162*, 443–455. https://doi.org/10.1016/j.neunet.2023.03.003

Qin, C., Yang, R., Huang, M., Liu, W., & Wang, Z. (2023). Spatial variation generation algorithm for motor imagery data augmentation: Increasing the density of sample vicinity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 3675–3686.

Qin, C., Yang, R., You, W., Chen, Z., Zhu, L., Huang, M., & Wang, Z. (2024). EEGUnity: Open-source tool in facilitating unified EEG datasets towards large-scale EEG model. *Preprint*. arXiv:2410.07196.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a general-purpose natural language processing task solver? *Preprint*. arXiv:2302.06476.

Qiu, C., Song, Y., Liu, Y., Zhu, Y., Han, K., Sheng, V. S., & Liu, Z. (2024). MMMViT: Multiscale multimodal vision transformer for brain tumor segmentation with missing modalities. *Biomedical Signal Processing and Control*, *90*, 105827. https://doi.org/10.1016/j.bspc.2023.105827

Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., & Bengio, Y. (2020). Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6989–6993).

Roche, J., De-Silva, V., Hook, J., Moencks, M., & Kondoz, A. (2021). A multimodal data processing system for LiDAR-based human activity recognition. *IEEE Transactions on Cybernetics*, *52*(10), 10027–10040. https://doi.org/10.1109/TCYB.2021.3085489

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* (pp. 766–775).

Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology*, *51*(5), 1450–1484. https://doi.org/10.1111/bjet.v51.5

Sharma, A., & Hamarneh, G. (2019). Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. *IEEE Transactions on Medical Imaging*, *39*(4), 1170–1183. https://doi.org/10.1109/TMI.42

Shen, Y., & Gao, M. (2019). Brain tumor segmentation on MRI with missing modalities. In *Information Processing in Medical Imaging: 26th International Conference* (pp. 417–428).

Shen, M., Zhang, H., Cao, Y., Yang, F., & Wen, Y. (2021). Missing data imputation for solar yield prediction using temporal multi-modal variational auto-encoder. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 2558–2566).

Shen, L., Zhu, W., Wang, X., Xing, L., Pauly, J. M., Turkbey, B., S. A. Harmon, Sanford, T. H., Mehralivand, S., Choyke, P. L., & B. J. Wood (2020). Multi-domain image completion for random missing input data. *IEEE Transactions on Medical Imaging*, *40*(4), 1113–1122. https://doi.org/10.1109/TMI.2020.3046444

Shi, Y., Paige, B., & Torr, P. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, *32*, 15692–15703.

Shi, J., Yu, L., Cheng, Q., Yang, X., Cheng, K.-T., & Yan, Z. (2023). M^{2}FTrans: Modality-masked fusion transformer for incomplete multi-modality brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, *28*(1), 379–390.

Soviany, P., Ionescu, R. T., Rota, P., & Sebe, N. (2022). Curriculum learning: A survey. *International Journal of Computer Vision*, *130*(6), 1526–1565. https://doi.org/10.1007/s11263-022-01611-x

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2022). From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 539–559. https://doi.org/10.1109/TPAMI.2022.3148210

Sun, Y., Liu, Z., Sheng, Q. Z., Chu, D., Yu, J., & Sun, H. (2024). Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, *110*, 102454. https://doi.org/10.1016/j.inffus.2024.102454

Sun, W., Ma, F., Li, Y., Huang, S.-L., Ni, S., & Zhang, L. (2021). Semi-supervised multimodal image translation for missing modality imputation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4320–4324).

Sun, J., Zhang, X., Han, S., Ruan, Y.-P., & Li, T. (2024). RedCore: Relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. in *Proceedings of the AAAI Conference on Artificial Intelligence* (vol. 38, no. 13, pp. 1573–15182).

Suo, Q., Zhong, W., Ma, F., Yuan, Y., Gao, J., & Zhang, A. (2019). Metric learning on healthcare data with incomplete modalities. In *IJCAI* (Vol. 3534, p. 3540).

Sutter, T., Daunhawer, I., & Vogt, J. (2020). Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in Neural Information Processing Systems*, *33*, 6100–6110.

Suzuki, M., & Matsuo, Y. (2022). A survey of multimodal deep generative models. *Advanced Robotics*, *36*(5–6), 261–278. https://doi.org/10.1080/01691864.2022.2035253

Tran, L., Liu, X., Zhou, J., & Jin, R. (2017). Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1405–1414).

Vadacchino, S., Mehta, R., Sepahvand, N. M., Nichyporuk, B., Clark, J. J., & Arbel, T. (2021). Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images. In *Medical Imaging with Deep Learning* (pp. 787–801).

Van Tulder, G., & de Bruijne, M. (2015). Why does synthesized data improve multi-sequence classification? In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference* (pp. 531–538).

Vapnik, V., & Izmailov, R. (2015). Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, *16*(1), 2023–2049.

Wan, Z., Yang, R., Huang, M., Zeng, N., & Liu, X. (2021). A review on transfer learning in EEG signal analysis. *Neurocomputing*, *421*, 1–14. https://doi.org/10.1016/j.neucom.2020.09.017

Wang, C., Butts, C. T., Hipp, J. R., Jose, R., & Lakon, C. M. (2016). Multiple imputation for missing edge data: A predictive evaluation method with application to add health. *Social Networks*, *45*, 89–98. https://doi.org/10.1016/j.socnet.2015.12.003

Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., & Carneiro, G. (2023). Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15878–15887).

Wang, X., Chen, Y., & Zhu, W. (2021). A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 4555–4576.

Wang, Q., Ding, Z., Tao, Z., Gao, Q., & Fu, Y. (2018). Partial multi-view clustering via consistent GAN. In *2018 IEEE International Conference on Data Mining* (pp. 1290–1295).

Wang, Q., Ding, Z., Tao, Z., Gao, Q., & Fu, Y. (2020). Generative partial multi-view clustering. *Preprint*. arXiv:2003.13088.

Wang, Y., Li, Y., & Cui, Z. (2024). Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, *36*, 17117–17128.

Wang, Q., Lian, H., Sun, G., Gao, Q., & Jiao, L. (2020). iCmSC: Incomplete cross-modal subspace clustering. *IEEE Transactions on Image Processing*, *30*, 305–317. https://doi.org/10.1109/TIP.83

Wang, H., Ma, C., Liu, Y., Chen, Y., Tian, Y., Avery, J., Hull, L., & Carneiro, G. (2024). Enhancing multi-modal learning: Meta-learned cross-modal knowledge distillation for handling missing modalities. *Preprint*. arXiv:2405.07155.

Wang, Q., Zhan, L., Thompson, P., & Zhou, J. (2020). Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1828–1838).

Wang, Y., Zhang, Y., Liu, Y., Lin, Z., Tian, J., Zhong, C., Shi, Z., Fan, J., & He, Z. (2021). ACN: Adversarial co-training network for brain tumor segmentation with missing modalities. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference* (pp. 410–420).

Wang, L., Zhang, X., Su, H., & Zhu, J. (2024). A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5362–5383.

Wang, Y., Zhou, L., Yu, B., Wang, L., Zu, C., Lalush, D. S., Lin, W., Wu, X., Zhou, J., & Shen, D. (2018). 3D auto-context-based locality adaptive multi-modality gans for PET synthesis. *IEEE Transactions on Medical Imaging*, *38*(6), 1328–1339. https://doi.org/10.1109/TMI.42

Wei, M., Huang, M., & Ni, J. (2023 Sep.). Cross-subject EEG channel selection method for lower limb brain-computer interface. *International Journal of Network Dynamics and Intelligence*, *2*(3), 100008. https://doi.org/10.53941/ijndi.2023.100008

Wu, R., Chen, X., Zhuang, Y., & Chen, B. (2020). Multimodal shape completion via conditional generative adversarial networks. In *Computer Vision–ECCV 2020: 16th European Conference* (pp. 281–296).

Wu, Z., Dadu, A., Tustison, N., Avants, B., Nalls, M., Sun, J., & Faghri, F. (2024). Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*.

Wu, M., & Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, *31*, 5580–5590.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., & Ye, J. (2013). Multi-source learning with block-wise missing data for Alzheimer's disease prediction. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 185–193).

Xiao, Y., Tian, Z., Yu, J., Zhang, Y., Liu, S., Du, S., & Lan, X. (2020). A review of object detection based on deep learning. *Multimedia Tools and Applications*, *79*(33–34), 23729–23791. https://doi.org/10.1007/s11042-020-08976-6

Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(10), 12113–12132. https://doi.org/10.1109/TPAMI.2023.3275156

Xue, Y., Yang, R., Chen, X., Liu, W., Wang, Z., & Liu, X. (2024). A review on transferability estimation in deep transfer learning. *IEEE Transactions on Artificial Intelligence*, *5*(12), 5894–5914.

Xue, Y., Yang, R., Chen, X., Song, B., & Wang, Z. (2024). Separable convolutional network-based fault diagnosis for high-speed train: A gossip strategy-based optimization approach. *IEEE Transactions on Industrial Informatics*, *21*(1), 307–316.

Yang, Q., Guo, X., Chen, Z., Woo, P. Y., & Yuan, Y. (2022). D^2-net: Dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, *41*(10), 2953–2964. https://doi.org/10.1109/TMI.2022.3175478

Yang, Y., Huang, C., Xia, L., & Li, C. (2022). Knowledge graph contrastive learning for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1434–1443).

Yang, H., Sun, J., & Xu, Z. (2023). Learning unified hyper-network for multi-modal MR image synthesis and tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, *42*(12), 3678–3689.

Yang, G., Tao, H., Wu, K., Du, R., & Zhong, Y. (2024). Fault diagnosis of harmonic drives using multimodal collaborative meta network with severely missing modality. *IEEE Transactions on Industrial Informatics*, *20*(8), 10366–10374.

Yang, C., Zhu, F., Liu, G., Han, J., & Hu, S. (2022). Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 4505–4514).

Ye, M., Shen, J., & Shao, L. (2020). Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, *16*, 728–739. https://doi.org/10.1109/TIFS.10206

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *Preprint*. arXiv:1702.01923.

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, *31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

Yu, S., Wang, J., Hussein, W., & Hung, P. C. (2024). Robust multimodal federated learning for incomplete modalities. *Computer Communications*, *214*, 234–243. https://doi.org/10.1016/j.comcom.2023.12.003

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., & Ye, J. (2012). Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1149–1157).

Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, *64*, 243–252. https://doi.org/10.1613/jair.1.11345

Zeng, J., Liu, T., & Zhou, J. (2022). Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1545–1554).

Zeng, J., Zhou, J., & Liu, T. (2022b). Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. *IEEE Transactions on Multimedia*, *25*, 6301–6314. https://doi.org/10.1109/TMM.2022.3207572

Zhan, B., Li, D., Wu, X., Zhou, J., & Wang, Y. (2021). Multi-modal MRI image synthesis via GAN with multi-scale gate mergence. *IEEE Journal of Biomedical and Health Informatics*, *26*(1), 17–26. https://doi.org/10.1109/JBHI.2021.3088866

Zhan, Y., & Yang, R. (2023). TransVAT: Transformer encoder with variational attention for few-shot fault diagnosis. In *2023 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS)* (pp. 1–6).

Zhan, Y., Yang, R., Zhang, Y., & Wang, Z. (2024). Mitigating catastrophic forgetting in cross-domain fault diagnosis: An unsupervised class incremental learning network approach. *IEEE Transactions on Instrumentation and Measurement*, *74*, 35–47.

Zhang, C., Chu, X., Ma, L., Zhu, Y., Wang, Y., Wang, J., & Zhao, J. (2022). M^{3}care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 2418–2428).

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, Z., Luo, H., Zhu, L., Lu, G., & Shen, H. T. (2022). Modality-invariant asymmetric networks for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, *35*(5), 5091–5104.

Zhang, S., & Metaxas, D. (2024). On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis*, *91*, 102996. https://doi.org/10.1016/j.media.2023.102996

Zhang, Y., Peng, C., Wang, Q., Song, D., Li, K., & Zhou, S. K. (2024). Unified multi-modal image synthesis for missing modality imputation. *IEEE Transactions on Medical Imaging*, *44*(1), 4–18.

Zhang, D., Wang, C., Chen, T., Chen, W., & Shen, Y. (2024). Scalable swin transformer network for brain tumor segmentation from incomplete MRI modalities. *Artificial Intelligence in Medicine*, *149*, 102788. https://doi.org/10.1016/j.artmed.2024.102788

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.

Zhang, W., Xu, D., Zhang, J., & Ouyang, W. (2021). Progressive modality cooperation for multi-modality domain adaptation. *IEEE Transactions on Image Processing*, *30*, 3293–3306. https://doi.org/10.1109/TIP.2021.3052083

Zhang, Y., Yang, J., Tian, J., Shi, Z., Zhong, C., Zhang, Y., & He, Z. (2021). Modality-aware mutual learning for multi-modal medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference* (pp. 589–599).

Zhang, Y., Zhang, Y., Guo, W., Cai, X., & Yuan, X. (2022). Learning disentangled representation for multimodal cross-domain sentiment analysis. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(10), 7956–7966. https://doi.org/10.1109/TNNLS.2022.3147546

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* (pp. 2608–2618).

Zheng, T., Li, A., Chen, Z., Wang, H., & Luo, J. (2023). Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking* (pp. 1–15).

Zhong, M., Zhu, X., Xue, T., & Zhang, L. (2023). An overview of recent advances in model-based event-triggered fault detection and estimation. *International Journal of Systems Science*, *54*(4), 929–943. https://doi.org/10.1080/00207721.2022.2146990

Zhou, T., Canu, S., Vera, P., & Ruan, S. (2021b). Latent correlation representation learning for brain tumor segmentation with missing MRI modalities. *IEEE Transactions on Image Processing*, *30*, 4263–4274. https://doi.org/10.1109/TIP.2021.3070752

Zhou, T., Fu, H., Chen, G., Shen, J., & Shao, L. (2020). Hi-net: Hybrid-fusion network for multi-modal MR image synthesis. *IEEE Transactions on Medical Imaging*, *39*(9), 2772–2781. https://doi.org/10.1109/TMI.42

Zhou, T., Ruan, S., & Hu, H. (2023). A literature survey of MR-based brain tumor segmentation with missing modalities. *Computerized Medical Imaging and Graphics*, *104*, 102167. https://doi.org/10.1016/j.compmedimag.2022.102167

Zhu, F., Zhang, X.-Y., & Liu, C.-L. (2023). Class incremental learning: A review and performance evaluation. *Acta Automatica Sinica*, *49*(3), 635–660.