Learning on Multimodal Graphs

Ciyuan Peng, *Federation University Australia, Ballarat, VIC, 3350, Australia*
Estrid He and Feng Xia, *RMIT University, Melbourne, VIC, 3000, Australia*

Multimodal data pervades various domains, such as health care, social media, and transportation, where multimodal graphs play a pivotal role. Multimodal graph learning (MGL) is essential for successful artificial intelligence applications, encompassing diverse graph types, modalities, techniques, and scenarios. This survey conducts a comparative analysis of existing works in MGL, elucidating how MGL is achieved across different graph types and exploring the characteristics of prevalent learning techniques. Unlike existing MGL surveys that focus on multimodal graphs (MGs), i.e., how to construct relational graphs from multimodal data, this article focuses on the learning techniques over MGs, i.e., the methodological designs for achieving effective feature extraction and integration over constructed MGs. Additionally, we delineate significant applications of MGL and offer insights into future directions in this domain. Consequently, this article serves as a foundational resource for researchers seeking to comprehend existing MGL techniques and their applicability across diverse scenarios.

Underpinned by significant advancements in artificial intelligence (AI) and machine learning techniques, there is a growing interest in multimodal learning, which involves comprehending and fusing knowledge from diverse modalities, such as texts, images, audios, and various other forms of data.[1] Compared with unimodal learning that relies on a single data source in one modality, multimodal learning integrates the knowledge from multiple modalities, and hence, allows the development of more effective, accurate, and robust machine learning models, such as large vision models.

As a special type of multimodal learning, multimodal graph learning (MGL) has become an emerging topic, due to the prevalence of multimodal graphs (MGs).[2] Numerous types of multimodal data are present in a graph format, usually referred to as MGs where nodes represent entities of heterogeneous types and edges indicate connections amongst them. Notable examples of MGs include health-care databases of patients' medical records (e.g., lab test results, clinical records, and imaging reports), social networks representing the interactions among Internet entities (e.g., online users, organizations, and commercial companies), and interconnected transport networks linking points of interest through different modes of transportation. Hence, developing effective learning techniques over multimodal graphs is regarded as a significant research opportunity, which will provide substantial benefits to downstream applications (e.g., data management, information retrieval, and outlier detection) encompassing various domains, such as biomedicine, health care, multimedia, and transportation.

MGL aims to leverage the relational representations of MGs to fully explore the intra- and inter-modal correlations of multimodal data. A key challenge in MGL involves effectively processing and fusing knowledge extracted from multiple modalities, given the complex graph topology. This is different from other types of multimodal learning where data have a clear and uniform structure. In MGL, data fusion must be guided by the graph structure. This is particularly challenging when the graph topology varies

Digital Object Identifier 10.1109/MIS.2026.3650817
Date of publication 12 January 2026; date of current version 12 April 2026.

across different modalities. Thus, a number of studies have been proposed for MGL, covering a wide range of tasks, including medical tasks (e.g., brain disease detection,[3]) development of conversation systems (e.g., visual question answering and conversation understanding,[4]) information retrieval (e.g., information extraction[5] and image retrieval,[6]) social media analysis (e.g., recommendation,[2]) and key knowledge graph (KG) tasks (e.g., KG completion,[7]) and so on. These works study different types of MGs, e.g., graphs where data modalities vary within nodes, across nodes, and even across multiple graphs. Meanwhile, they approach the problems of MGL using different techniques, choosing various types of graph learning architectures to achieve effective data fusion in their particular application contexts. The rich diversities of the existing works on MGL call for a comprehensive survey, which systematically summarizes the progress in this domain so far.

To the best of our knowledge, there is only one survey paper on multimodal graph learning,[8] which mainly discusses the potential of leveraging the characteristics of graphs to achieve multimodal learning. It presents how MGs can be constructed in various application scenarios, based on which graph learning technology can be applied. However, it does not discuss the intricacy in the design of multimodal graph learning techniques, i.e., how to choose the most appropriate model to process a certain type of MGs. In this survey, we fill this gap and provide a systematic overview of existing MGL studies in terms of their methodological designs. Particularly, we present an extensive summary outlining the *strengths/limitations* of popular and elaborately chosen deep learning architectures for various types of MGs.

In what follows, we formally define MGs. Then we present a taxonomy of existing MGL techniques, and present a comprehensive comparative overview of these techniques. We summarize vital applications of MGL and list important pointers to useful resources for implementing relevant techniques for these applications. Finally, we discuss the remaining challenges and our insights into future directions.

MULTIMODAL GRAPHS

In this article, we define MGs as graphs carrying data in heterogeneous modalities, such as a combination of visual, textual, and acoustic data.[9] We focus on a popular setting where nodes carry multimodal data while the features of edges are unimodal and reflect the connections of nodes. As such, based on the distribution of different data modalities across nodes, we classify MGs into three types: feature-level MGs, node-level MGs, and graph-level MGs.

Figure 1 illustrates the three types of MGs:

- *Feature-level MGs*: These are graphs where each node stores multimodal features. Feature-level MGs are commonly used for a set of interconnected entities, each of which has descriptors of different modalities, e.g., a multimodal conversation system.[4]
- *Node-level MGs*: These are graphs where each node carries unimodal features while the feature modalities vary across nodes. Node-level MGs are constructed to represent connections amongst entities that are described in different modalities, e.g., a multimodal graph that connects images to their textual keywords.[10]
- *Graph-level MGs*: These are graphs that contain multiple subgraphs, each of which stores features of a sole modality. In graph-level MGs, the nodes in subgraphs usually correspond to the same set of entities, while their connections vary with different modalities.[2]

A TAXONOMY OF MGL

Based on the approaches for processing MGs, we categorize existing MGL works into three predominant types: 1) multimodal graph convolution network (MGCN), 2) multimodal graph attention network (MGAT), and 3) multimodal graph contrastive learning (MGCL). Note that there exist works that use a combination of different learning models for different purposes (e.g., encoders for different modalities). Hence, we categorize all works by their techniques used for the purpose of *data fusion*. Figure 2 illustrates three types of MGL. Table 1 presents a list of the representative MGL methods.

Multimodal Graph Convolutional Network

Graph convolutional network (GCN) is a graph model based on the convolutional neural network that updates node representations through a convolution operation. GCN primarily focuses on capturing the relations between nodes and their neighbors, propagating and aggregating information across nodes using adjacency matrices. Recently, to apply GCN over MGs, MGCN has been proposed.[3,6] Notably, due to the advantage of MGCN in exploring the relations amongst nodes, it is highly effective in revealing the correlations between different modalities when applied on

node-level MGs.[10] Hence, MGCN shows great potential in handling node-level MGs, compared to the other two types of MGs, in terms of extracting cross-modal relations.

Cross-Modal Relation Modeling for Node-Level MGs

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a node-level multimodal graph, where $\mathcal{V}$ and $\mathcal{E}$ represent the node and edge sets, and let M be the set of modalities contained in $\mathcal{G}$. We denote the i-th node in $\mathcal{V}$ as $v_i^{m_1} \in \mathcal{V}$ where $m_1 \in M$ is its modality. We use $\mathbf{X} \in \mathbb{R}^{n \times d}$ to represent the node feature matrix (e.g., $\mathbf{x}_i^{m_1} \in \mathbb{R}^d$ represents the feature vector of node $v_i^{m_1}$), where n is the number of nodes and d is the feature dimension. A multimodal adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be obtained. Let $N(v_i^{m_1})$ be the set of neighboring nodes of $v_i^{m_1}$ and $v_j^{m_2} \in N(v_i^{m_1})$ be one neighbor of $v_i^{m_1}$. The adjacency feature $a_{ij}^{(m_1,m_2)} \in \mathbf{A}$ can quantify the correlation between modalities m_1 and m_2. Feeding $\mathcal{G}$ into MGCN, the output of l-th layer of MGCN is

$$\mathbf{H}^{(l+1)} = \sigma(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}). \quad (1)$$

Here, $\mathbf{H}^{(l+1)}$ represents the updated node feature matrix (output of l-th layer and input of $(l+1)$-th layer), and $\mathbf{H}^{(1)} = \mathbf{X}$. The function $\sigma(\cdot)$ represents the activation function, $\hat{\mathbf{D}}$ represents the degree matrix of $\mathcal{G}$. $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, here $\mathbf{I}$ is the identity matrix, and $\mathbf{W}^{(l)}$ stands as the weight matrix. By updating the weight of $\hat{\mathbf{A}}$, MGCN can capture the relations among nodes, and further, extract the cross-modal relations.

In addition to node-level MGs, MGCN has also been applied to other MG types. For instance, Song et al. proposed a multichannel pooling GCN to encode feature-level multimodal brain graphs for brain disease detection.[3] However, MGCN shows more significant power for node-level MGs.

Characteristics of MGCN

- *MGCN is effective in learning complex interactions amongst heterogeneous nodes*: In node-level MGs, neighboring nodes usually have heterogeneous modalities. By processing the multimodal adjacency matrix of a graph, MGCN can propagate the multimodal information residing in one node to its neighbors through convolution kernels, and thus, extract cross-modal relations amongst nodes to achieve effective data fusion.[10] This makes MGCN especially effective when there exist complex interactions amongst the heterogeneous nodes in a graph. For example, MGCN can effectively model the complex interactions in node-level multimodal knowledge graphs, wherein the connections among entities reflect diverse hyper-semantic interactions and cross-modal information (e.g., text-image relations).[6] In such graphs, MGCN explicitly considers the graph topology when aggregating heterogeneous information, and hence, facilitates a more profound understanding of the complicated interactions among multimodal nodes in a graph. Several existing works have demonstrated the superiority of MGCNs in learning effective representations for node-level MGs and achieving enhanced model performance in downstream tasks.[10]
- *MGCN is more effective in short-range information aggregation*: The reach of the convolution operation in MGCNs is limited by its filter size. As such, each node is primarily influenced by its neighboring nodes, and the propagation of information diminishes gradually with the increasing distance. Consequently, works have shown that MGCNs are restricted in capturing long-range information, and therefore, have relatively limited capability to process large-scale MGs.[11]
- *MGCN is class/label-driven, and does not explicitly guide a model to extract unique features of various modalities*: The strength of MGCN lies

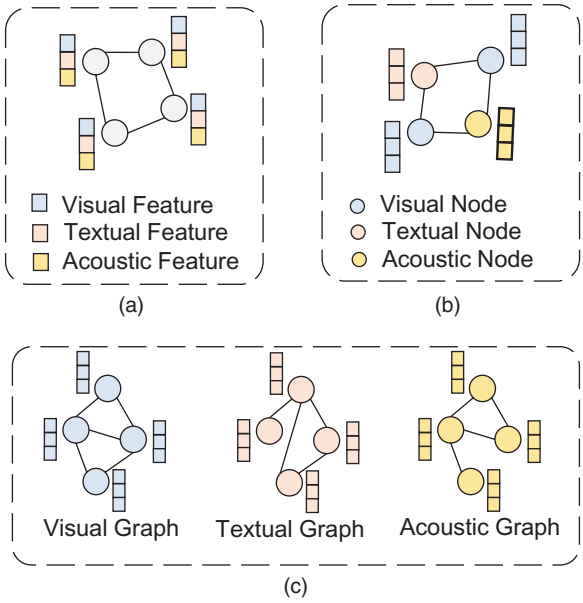

FIGURE 1. Three different types of MGs. (a) Feature-level MGs. (b) Node-level MGs. (c) Graph-level MGs.

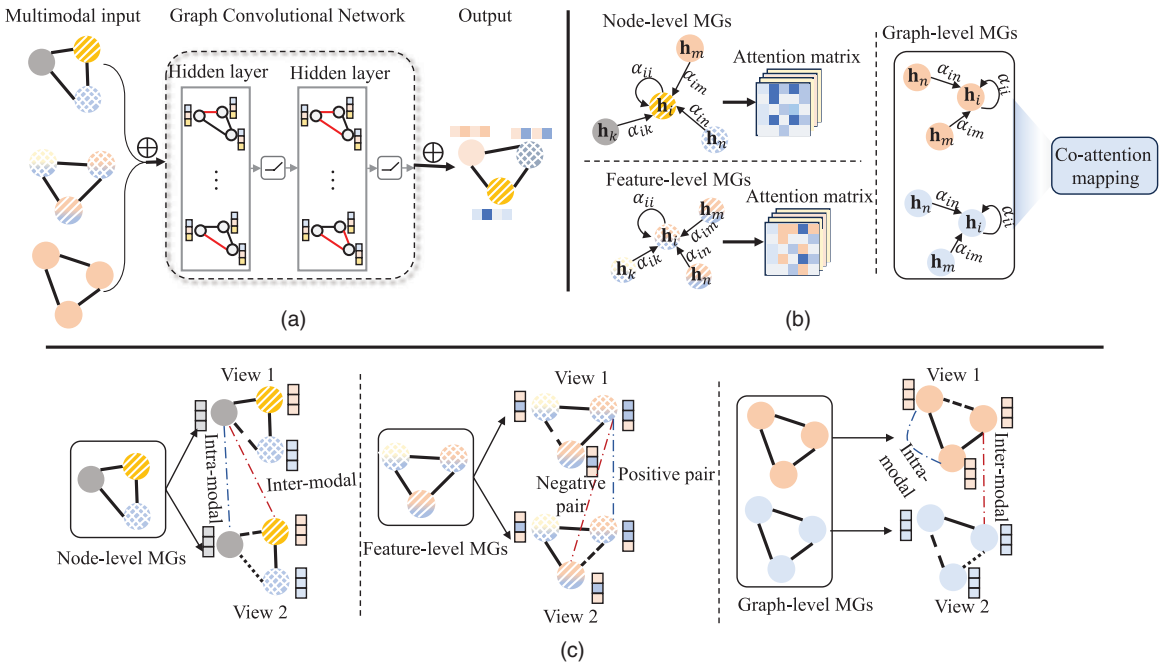

FIGURE 2. Three types of MGL. ⊕ symbol in MGCN indicates fusion operation, which can be applied either before learning or after learning. (a) MGCN. (b) MGAT. (c) MGCL.

in topology-guided feature fusion. However, since the feature fusion is primarily guided by supervision signals from class labels, it does not explicitly guide a model to identify unique characteristics of different modalities in MGs. Consequently, a trained model may not effectively identify the similarities/differences within and across modalities.[2] Therefore, several works have proposed to leverage contrastive learning to provide more comprehensive supervision signals, which will be detailed later.

Multimodal Graph Attention Network (GAT)

GAT utilizes the attention mechanism to weight the propagated information, dynamically measuring the importance of nodes in representation learning. Compared to GCN, GAT can flexibly learn the different attention weights of node neighbors and more effectively capture long-range information. Especially, when dealing with large-scale graphs, GAT can adaptively focus on important nodes by assigning them high attention weights, therefore achieving efficient learning. There are various GAT types according to different attention mechanisms, such as self-attention. MGAT extends GAT to the multimodal settings, integrating information from different modalities.[1]

Attention-Based Multimodal Information Fusion

MGAT implements multimodal information fusion by assigning different attention weights to nodes. For a given MG with $|\mathcal{V}|$ nodes, the latent representation of node v_i is calculated as

$$\mathbf{h}_i = \delta\Big(\sum_{v_j \in N(v_i)} \alpha_{ij} \mathbf{W}\mathbf{h}_j\Big), \tag{2}$$

where $\delta(\cdot)$ is the activation function (e.g., LeakyReLU$(\cdot)$), $N(v_i)$ represents the neighbor set of node v_i, and $\mathbf{W}$ is a learnable weight matrix. α_{ij} stands as the attention weight, which is calculated as

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{v_k \in N(v_i)} exp(e_{ik})}. \tag{3}$$

Here, e_{ij} is the relation weight of nodes v_i and v_j. Equation (2) is a standard formulation of MGAT, and can be optimized according to the setting of attention mechanisms. For node-level and feature-level MGs, MGAT integrates node features to obtain multimodal graph representations, denoted as $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{|\mathcal{V}|}]^\top$. Here, for feature-level MGs, multimodal features are expressed in each node vector. For example, the attributes of $\mathbf{h}_i$ are multimodal. For

node-level MGs, the attributes of each node are unimodal, while the modalities across nodes are different, denoted as $\mathbf{H} = [\mathbf{h}_1^{m_1}, \mathbf{h}_2^{m_2}, ..., \mathbf{h}_{|\mathcal{V}|}^{m_{|\mathcal{V}|}}]^\top$, here m_i indicates the modality of node v_i. If nodes v_i and v_j have the same modality, their corresponding m_i and m_j will be the same. In particular, for a node-level MG, because each node is assigned different attention weights, the overall contribution of each modality to the final graph representation will be different. Therefore, a tradeoff parameter has been suggested to balance the contribution of multiple modalities.[1]

Different from node-level and feature-level MGs, for graph-level MGs, MGAT first separately learns on each unimodal graph, and then utilizes a co-attention mapping to integrate graphs from different modalities, obtaining multimodal graph representations.[7] Formally, given a set of graph-level MGs containing M unimodal graphs, the representations of all the unimodal graphs can be gained by (2), denoted as $\{\mathbf{H}^1, \mathbf{H}^2, \ldots, \mathbf{H}^M\}$. Then, a co-attention mapping function $f(\cdot)$ is applied to obtain a multimodal graph representation. In literature, various methods have been proposed to construct the co-attention mapping function, such as semantic similarity measuring and gated attention aggregation.[7]

Characteristics of MGAT

- *MGAT may provide higher training efficiency*: MGAT captures node-level correlations via an attention mechanism, dynamically assigning attention weights to various nodes in MGs. Hence, it can potentially lead to more efficient training with reduced computational cost, especially when nodes have varying degrees of importance. As demonstrated by prior studies, attention-based information fusion models have fast training speed and superior performance when dealing with large heterogeneous graphs.[14] This indicates the ability of MGAT to perform efficient model training.
- *MGAT can capture long-range inter-modal dependencies*: MGAT excels in aggregating multimodal information globally over MGs. This is because the information fusion in MGAT is performed through an attention mechanism, which has long-range reach. Many works have proven the attention mechanism empowers each node with a global perception, allowing it to consider information from all other nodes in the heterogeneous graph rather than being confined to a local neighborhood. As such, MGAT is particularly advantageous when handling large-scale MGs.
- *MGAT may form biased multimodal representations for node-level MGs*: MGAT treats nodes differently by assigning them different importance scores. Hence, in node-level MGs, nodes from each modality will have different overall impacts on the final learned multimodal representations. This may lead the model to mistakenly consider certain modalities as less important, subsequently ignoring the information from these modalities and resulting in biased multimodal representations. Referring to one existing work,[1] which introduces a trainable bias to guide the information flow, a potential solution is introducing a tradeoff parameter to implement a flexible attention mechanism and balance the contribution of each modality.
- *MGAT is less robust when learning on noise-prone modalities*: Due to the noise-sensitive nature of the attention mechanism, MGAT is susceptible to irrelevant information from some noise-prone modalities (such as images and audios). Ignoring the noise accompanying irrelevant information may lead to modality contradiction.[7] Therefore, the weakness of the MGAT in handling noise may result in biased multimodal representations, especially when dealing with noise-prone modalities.

Multimodal Graph Contrastive Learning

Graph contrastive learning (GCL) focuses on learning distinct node representations by comparing positive samples (similar nodes) and negative samples (dissimilar nodes). Conceptually, GCL aims to maximize the feature consistency of similar nodes and minimize that of dissimilar nodes under different augmented graph views. Benefiting from the efficacy of GCL in enhancing the semantics of node representations and distinguishing different nodes, multimodal graph contrastive learning (MGCL) applies GCL to MGs.[15]

Intermodal and Intramodal Difference Extraction

Unlike MGCN and MGAT, MGCL explicitly emphasizes on extracting inter- and intramodal similarities/differences from MGs. It generates different graph views from different perspectives (e.g., modalities) and then applies the contrastive learning strategy to those graph views to learn distinguishable node representations. Given an MG and the latent representation of an

TABLE 1. A summary of different MGL methods.

Category	Method	MG Type	Modality	Task	Description	Advantage	Limitation
MGCN	[11]	N	*	Modality prediction and matching	Cell-feature graphs	Capturing high-order structural information	Limited to bipartite graphs
	[4]	G	T + V + A	Emotion recognition	Conversational interactions	Efficient learning on lengthy conversations	Lack of higher-order relationship modeling
	[10]	N	V + P	Freezing of gait (FoG) detection	FoG graphs	Adaptive to missing modalities	Limited hierarchical graph exploration
	[3]	F	SI + FI + Non-I	Brain disease detection	Brain graphs	Dual-modality fused brain connectivity network	High complexity
	[6]	N	T + V	Image retrieval	KGs	Multimodal knowledge hypergraphs	Lack of modality-specific structure
MGAT	[7]	G	T + V	Knowledge graph completion	KGs	Multilevel fusion	Less robust on noise-prone modalities
	[12]	N	T + V + A	Emotion recognition	Conversational interactions	Mitigating prediction biases	Limited learning capability due to simple Semantic graph learning
	[1]	G	T + V	Visual question answering	Semantic graphs	Structured graph-guided Transformer training	Potential biases from graph integration
	[13]	F	T + V	Analogical reasoning	KGs	Structure-based multimodal analogical reasoning	Not adaptable to unknown entities
	[14]	N	T + V	Knowledge graph completion	KGs	Structural information-based fusion	Potential bias in multimodal representations
MGCL	[15]	F	T + V + A	Emotion recognition	Conversational interactions	Fusion between global contextual and unimodal specific features	Highly dependent on data quality
	[5]	F	T + V	Information extraction	Document graphs	Centralized multimodal graph contrastive learning	Limited few-shot capability
	[2]	G	T + V	Recommendation	User-Item graphs	Reduction of multimodal noise contamination	Limited to short-range graph structures

(continued)

TABLE 1. A summary of different MGL methods. (*continued*)

Category	Method	MG Type	Modality	Task	Description	Advantage	Limitation
Others	[16]	F	T + V	Knowledge graph completion	KGs	Addressing spatial heterogeneity	Lack of cross-modal interactions
	[17]	G	SI + FI	Brain disease detection	Brain graphs	Hypergraph-regularized multimodal learning	Limited interpretability
	[18]	G	*	Medical diagnosis	Biomedical graphs	Omni-modal biomedical representation	Reliance on synthetic clinical text descriptions
	[19]	N	*	Biomedical KGs	Retrieval augmented generation	Multimodal knowledge graph-based RAG	Limited versatility
	[20]	F	SI + FI	Brain disease detection	Brain graphs	Multimodal graph prompt learning	Inability to preserve modality-specific structures

"F" means feature-level MG type, "N" is node-level MG type and "G" indicates graph-level MG type. "T" indicates text, "V" is vision, "A" is audio, "P" stands as pressure sequence, "SI" indicates structural neuroimaging phenotypes, "FI" is functional neuroimaging phenotypes, and "Non-I" means nonimaging data. Particularly, "*" indicates there are various modalities (such as cell, gene expression, protein, molecular, and so on) for multimodal biomedical graphs.

anchor sample $\mathbf{x}_i$, contrastive learning aims to minimize the following loss:

$$\mathcal{L} = -\log \frac{exp\left(h\left(\mathbf{x}_i, \mathbf{x}_i^{\text{pos}}\right)/\tau\right)}{\sum_{k=1}^{K} exp\left(h\left(\mathbf{x}_i, \mathbf{x}_k\right)/\tau\right)}. \tag{4}$$

Here, $h(\cdot)$ is a score function that measures the similarity between two sample representations. $\mathbf{x}_i^{\text{pos}}$ is the representation of a positive sample (similar node) with respect to the anchor node, and $\mathbf{x}_k$ indicates a negative sample. K and τ represent the number of negative samples and the temperature parameter, respectively.

For graph-level MGs, MGCL can generate different graph views according to modalities. Given a graph-level MG with M unimodal subgraphs, each unimodal graph is encoded by a graph neural network separately, resulting in M unimodal graph views $\{\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^M\}$. Here, a unimodal graph view is denoted as $\mathcal{G}^{m_1} = (\mathbf{X}^{m_1}, \mathbf{A}^{m_1})$, where $\mathbf{X}^{m_1}$ indicates the node representation of graph in modality m_1, and $\mathbf{A}^{m_1}$ is the adjacency matrix. Then, (4) can be applied to capture the similarity and dissimilarity of nodes across different modalities, thereby capturing inter-modal differences and similarities.[2] On the other hand, for feature-level and node-level MGs, MGCL obtains two distinct MG views by performing some graph augmentation methods, such as different transformations of MG structure.[5] Then, (4) is leveraged to distinguish node representations and capture inter- and intramodal differences. Notably, the tradeoff of intermodal and intramodal difference extraction can be reflected by the design of positive and negative samples. For example, setting nodes from different modalities as negative samples can capture intermodal differences, while setting negative samples within the same modality can explore intramodal differences.

Characteristics of MGCL

- *MGCL allows a model to learn correlations/dissimilarities amongst various modalities*: As a self-supervised learning strategy, MGCL does not require any manual labels from training data. This leads to an obvious advantage of MGCL: Lifting the requirement of data annotation and the vast amount of unlabeled data can be utilized for training. More importantly, MGCL generates effective supervision signals through contrasting training samples, which explicitly guides a model to learn the unique characteristics of each modality and the intermodal correlations. Consequently, this allows a model to link highly correlated modalities and distinguish differences

across modalities, forming more comprehensive multimodal representations. Several studies have shown that incorporating contrastive loss into the traditional class-driven loss function can improve the model performance significantly.[15]

- *Extending MGCL to graph-level MGs with more than two modalities remains a challenge*: Although MGCL can capture intermodal similarities and differences when applying it to graph-level MGs, the contrastive loss function is usually implemented on two graph views.[2] How to efficiently extend MGCL to graph-level MGs with more than two modalities is still an unexplored question in this area. When there are more than two modalities, (4) is no longer suitable. Thus, an advanced loss function is required to capture the relations between *each* pair of modalities, which may make the training process more complex.
- *The effectiveness of MGCL depends on the design of positive and negative samples*: A critical component in MGCL is to define positive/negative samples in (4). These samples have significant impact on the quality of learning results. Particularly, the design of positive/negative samples can influence intermodal and intra-modal difference extraction in MGs. As such, users of MGCL need to carefully define positive/negative samples given their application scenarios.[5]

Other Methods

In addition to the aforementioned three representative categories of methods, there are also other methods for processing MGs. We summarize them in this section. For example, optimal transport knowledge graph embeddings a multimodal knowledge graph embedding method, leverages multimodal knowledge (visual and textual) for the representation of entities and relations.[16] Moreover, Wang et al. presented hypergraph-regularized multimodal learning by graph diffusion for brain disorder diagnosis.[17] They proposed a multimodal phenotypic graph diffusion method to integrate multimodal brain graphs. Peng et al.[20] proposed multimodal graph prompt learning to integrate brain structural and functional graphs, achieving brain disease detection.

MG APPLICATIONS AND LIBRARIES

MGs for Knowledge Representations

Multimodal knowledge graphs (MKGs), one of the most representative applications of MGs, have attracted extensive attention in related domains such as natural language processing and computer vision, e.g., developing multimodal large language models (MLLMs) for construction of MKGs.[19,21] MKGs organize multimodal facts with entities and their relations from multiple sources, such as text, image, and video, comprehensively describing various knowledge with significantly improved quality.[7,14] As groundbreaking advancements in information expression, MKGs have been applied in various downstream applications, including dialogue systems, recommender systems, and information retrieval systems. These applications leverage MKGs' ability to integrate diverse modalities to enhance their performance and effectiveness.

Recently, MKG-related tasks, such as multimodal named entity recognition (MNER), multimodal link prediction (MLP), multimodal relation extraction (MRE) and multimodal knowledge reasoning (MKR), have emerged as the primary tasks of knowledge graph embeddings. More recently, along with the rise of MLLMs, many studies integrate large language models (LLMs) with MKGs for various tasks. For example, Li et al.[22] proposed an LLM-based approach to identify information from multimodal brand knowledge graphs, achieving phishing detection. Moreover, many domain-specific MKGs have been proposed to achieve multimodal relation modeling in different areas.[16] For example, medical MKGs have shown great potential in precise drug event prediction by integrating multimodal drug-drug interactions.

Libraries

To facilitate the research and applications of MKGs, many libraries/datasets have been made available including WN18-IMG, WN9-IMG, and FB15K-237-IMG, in which each entity has 10 images.[7] Some other MKG datasets are also widely used, such as MMDialKB[23] building upon human–human dialogues, MARS[13] for multimodal analogy reasoning tasks.

MGs for Biomedical Studies

Multimodal biomedical graphs (MBioGs) serve as powerful tools for organizing diverse biomedical information across different dimensions. They store comprehensive knowledge of biological systems, providing a holistic view that encompasses both molecular and cellular levels. At the biological molecular level, multimodal molecular graphs (MMGs) transcend traditional atom-centric representations, capturing the stereochemical structures of molecules along with their diverse molecular paraphrases. MMGs can depict intra- and intermolecular properties from multiple

scales, modeling molecule reactions and applied to artificial intelligence-informed medical analysis, such as drug discovery and molecular optimization.

At the biological cell level, multimodal single-cell graphs (MSGs) leverage multiscale interactions within single cells, such as protein–protein interactions and gene expressions, to measure comprehensive biological information. Learning on MSGs provides feature representations of key biomedical entities (e.g., proteins, drugs, and genes), capturing high-order relations and significantly benefiting single-cell analysis.[11] MBioGs contribute to advancing our understanding of complex biological systems by integrating multimodal data sources.

Many studies have explored the integration of LLMs and large vision models (LVMs) with biomedical data to advance various applications, such as constructing and augmenting MBioGs. For example, Gtp-4o[18] leverages an LLM, such as MiniGPT-4, to generate synthetic text descriptions for pathology slides, thereby enhancing the representations of MBioGs. Moreover, KG-RAG[19] integrates multimodal biomedical knowledge graphs with a pretrained LLM in a retrieval augmented generation (RAG) framework for biomedical text generation.

Libraries

Some open biomedical libraries provide available datasets for MBioG analysis. Recently, a multimodal single-cell dataset[24] has been released for multimodal single-cell integration analysis. In addition, various other biomedical corpora (e.g., GENIA corpus[25] and BioNLP Shared Task[26]) can be utilized to construct MBioGs.

MGs for Brain Science

Various neuroimaging techniques, including electrogastrography (EGG), magnetic resonance imaging (MRI), and diffusion tensor imaging (DTI), are developed to collect multimodal neuroimaging data for brain health analysis.[3] In particular, to explore the complex information interaction mechanism among brain regions, multimodal brain graphs (MBrainGs) are proposed to integrate multimodal neuroimaging data and depict various brain structural and functional connectivities. Many studies have shown the superiority of MBrainGs in capturing multidimensional relations of brain regions. Notably, LLMs and LVMs have shown significant potential in advancing MBrainG construction. For example, MMGPL[20] leverages GPT-4 to obtain disease concepts to enhance MBrainGs.

According to the characteristics of neuroimaging data, MBrainGs can be categorized into several different types. First, multimodal brain functional graphs (Mbrain-FGs) are constructed by exploring brain functional connectivities based on multimodal functional neuroimaging data [e.g., magnetoencephalography (MEG) and EEG]. Since they capture diverse functional information, Mbrain-FGs can be employed for brain activity analysis. Second, multimodal brain structural graphs (Mbrain-SGs) fuse structural information from multimodal structural neuroimaging data (e.g., structural MRI and DTI) to model brain structural connectivities. Third, multimodal brain structural-functional graphs (Mbrain-SFGs) integrate both structural and functional data to achieve joint representations of brain structure and function, showing great potential in advancing various applications, especially precise brain disorder diagnosis.[17] In addition, some works also embed features extracted from nonimaging data (e.g., demographic data) into MBrainGs to enhance the graph.[9] Learning on MBrainGs provides researchers with a more comprehensive understanding of the brain structure and function, facilitating the in-depth exploration of multidimensional relations within the brain.

Libraries

There exist multiple brain health datasets that are publicly available. ADNI[17] and OASIS-3[9] datasets collect the multimodal neuroimaging data of subjects from different age groups for studies on Alzheimer's Disease. ABIDE[9] dataset focuses on autism spectrum disorders and collects multimodal neuroimaging data from more than thousands of subjects.

CHALLENGES AND OUTLOOKS

Data Imbalance Across Modalities

In real-life applications, the collected multimodal dataset may have missing or corrupted data due to various reasons, e.g., faulty sensors, database failures, and data security issues.[27] This may result in data imbalance across modalities if one or more modalities are significantly damaged and become minority modalities. Whether current MGL models can effectively handle such issues is a largely unexplored topic. As such, comprehensive studies that evaluate the reliability and trustworthiness of existing MGL models under moderate to significant modality imbalance will be highly valuable to this field. Furthermore, remediation strategies that may alleviate data imbalance issues, such as MG augmentation and data resampling,

are worth in-depth investigation so as to improve the robustness of MGL models in such a context.

Trustworthy Multimodal Alignment

To accurately integrate the knowledge provided by different data modalities, it is crucial to align the information across modalities by mapping their representations into a shared space where both intramodal consistency and intermodal complementarity are preserved. We suggest that current multimodal alignment methods can be improved mainly from two aspects. First, more explicit guidance on how to match relevant information from multiple modalities is needed, e.g., locating the correct image from a document given a textual description of the image in the same document. In many current vision–language models, the multimodal matching is done in a brute-force manner without explicit guidance.[28] We argue that such multimodal matching is less useful for MGs, especially graph-level MGs, due to the heterogeneous topology of graphs. Therefore, developing graph structure-adaptive multimodal matching methods is needed. Second, the reliability of each modality needs to be explicitly considered during multimodal alignment. In real-life applications, due to random noises, it is common for data modalities to have different accuracies. Thus, we believe that further methodological improvement from this aspect will enhance the robustness of the multimodal alignment.

Temporal Multimodal Graph Learning

An MG may evolve over time with changes in graph topology and node/edge features. Learning on a temporally-evolving MG is challenging, due to the dynamic interactions amongst data modalities. Performing dynamic data fusion is a much more challenging task. Compared to the learning on static MGs, learning over temporal MGs requires the extraction of higher-order features, i.e., the spatiotemporal correlations amongst nodes.[29] This challenges the learning capability of MGL models, especially on how to accurately align the information across different modalities on both the temporal and spatial dimensions. In the future, more advanced techniques, such as the adaptation of continual learning, are worth being studied for this field.

Computational Efficiency of Large-Scale MGs

With the exponentially growing sizes of unimodal models (e.g., 110M parameters for BERT and 175B parameters for GPT3), the computational complexity of many MGL models is also rapidly growing, due to the need to process different modalities. This poses significant challenges in the training of these models and also the deployment of these models onto resource-constrained devices, such as laptops and mobile phones at user ends.[30] Hence, improving the computational efficiency of MGL models will be of significant value to this field. It has been pointed out that various existing multimodal models, e.g., DALL-E and CM3, rely on knowledge memorization, and thus, need to grow model parameters when learning large-scale datasets. This calls for techniques, which guide a model to learn high-quality and generalizable features from data rather than simply memorizing and storing features. Other techniques, which deal with model sizes directly, such as model compression and distributed learning will also be valuable to study.

CONCLUSION

MGL is an emerging field, which aims to leverage graph learning techniques to achieve effective data fusion and representation learning over multimodal data. In this article, we presented an overview of these techniques for representation learning on multimodal graphs. We reviewed how machine learning techniques are designed for various types of multimodal graphs to achieve effective data fusion. We discussed the characteristics (e.g., advantages and drawbacks) of those popular designs on their effectiveness in mining the intra- and intermodal correlations. We present unique perspectives on MG typology and provide practical guidance on selecting appropriate learning strategies for different MG types. In addition, we summarized the key applications of multimodal graph learning with crucial pointers to implementation resources. This article provides a comprehensive discussion on the current progress and remaining challenges in this domain. Based on these discussions, we presented our future outlooks of emerging topics in this domain, including issues around data quality, trustworthiness, dynamic evolving, and computational efficiency. This article can be used as a guideline for new researchers.

REFERENCES

1. X. He and X. Wang, "Multimodal graph transformer for multimodal question answering," in *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2023, pp. 189–200.
2. K. Liu, F. Xue, D. Guo, P. Sun, S. Qian, and R. Hong, "Multimodal graph contrastive learning for multimedia-based recommendation," *IEEE Trans. Multimedia*, vol. 25, pp. 9343–9355, 2023, doi: 10.1109/TMM.2023.3251108.

3. X. Song et al., "Multicenter and multichannel pooling GCN for early AD diagnosis based on dual-modality fused brain network," *IEEE Trans. Med. Imag.*, vol. 42, no. 2, pp. 354–367, Feb. 2023, doi: 10.1109/TMI.2022.3187141.
4. M. Li et al., "FrameERC: Framelet transform based multimodal graph neural networks for emotion recognition in conversation," *Pattern Recognit.*, vol. 161, May 2025, Art. no. 111340, doi: 10.1016/j.patcog.2024.111340.
5. C.-Y. Lee et al., "FormNetV2: Multimodal graph contrastive learning for form document information extraction," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers)*, 2023, pp. 9011–9026.
6. Y. Zeng, Q. Jin, T. Bao, and W. Li, "Multi-modal knowledge hypergraph for diverse image retrieval," *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 3, 3376–3383, 2023, doi: 10.1609/aaai.v37i3.25445.
7. X. Chen et al., "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2022, pp. 904–915, doi: 10.1145/3477495.3531992.
8. Y. Ektefaie, G. Dasoulas, A. Noori, M. Farhat, and M. Zitnik, "Multimodal learning with graphs," *Nature Mach. Intell.*, vol. 5, no. 4, pp. 340–350, 2023, doi: 10.1038/s42256-023-00624-6.
9. S. Kim, N. Lee, J. Lee, D. Hyun, and C. Park, "Heterogeneous graph learning for multi-modal medical data analysis," *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 4, 5141–5150, 2023, doi: 10.1609/aaai.v37i4.25643.
10. K. Hu et al., "Graph fusion network-based multimodal learning for freezing of gait detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 3, pp. 1588–1600, Mar. 2023, doi: 10.1109/TNNLS.2021.3105602.
11. H. Wen, J. Ding, W. Jin, Y. Wang, Y. Xie, and J. Tang, "Graph neural networks for multimodal single-cell data integration," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining (KDD)*, 2022, pp. 4153–4163, doi: 10.1145/3534678.3539213.
12. J. Shi, M. Li, Y. Chen, L. Cui, and L. Bai, "Multimodal graph learning with framelet-based stochastic configuration networks for emotion recognition in conversation," *Inf. Sci.*, vol. 686, Jan. 2025, Art. no. 121393, doi: 10.1016/j.ins.2024.121393.
13. N. Zhang, L. Li, X. Chen, X. Liang, S. Deng, and H. Chen, "Multimodal analogical reasoning over knowledge graphs," 2023, *arXiv:2210.00312*.
14. B. Shang, Y. Zhao, J. Liu, and D. Wang, "LAFA: Multimodal knowledge graph completion with link aware fusion and aggregation," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 8, 8957–8965, 2024, doi: 10.1609/aaai.v38i8.28744.
15. D. Li, Y. Wang, K. Funakoshi, and M. Okumura, "Joyful: Joint modality fusion and graph contrastive learning for multimodal emotion recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 16,051–16,069.
16. Z. Cao, Q. Xu, Z. Yang, Y. He, X. Cao, and Q. Huang, "OTKGE: Multi-modal knowledge graph embeddings via optimal transport," in *Proc. 36th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2022, pp. 39,090–39,102.
17. M. Wang, W. Shao, S. Huang, and D. Zhang, "Hypergraph-regularized multimodal learning by graph diffusion for imaging genetics based Alzheimer's disease diagnosis," *Med. Image Anal.*, vol. 89, Oct. 2023, Art. no. 102883, doi: 10.1016/j.media.2023.102883.
18. C. Li et al., "GTP-4o: Modality-prompted heterogeneous graph learning for omni-modal biomedical representation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland: Springer-Verlag, 2025, pp. 168–187.
19. K. Soman et al., "Biomedical knowledge graph-optimized prompt generation for large language models," *Bioinformatics*, vol. 40, no. 9, 2024, Art. no. btae560, doi: 10.1093/bioinformatics/btae560.
20. L. Peng, S. Cai, Z. Wu, H. Shang, X. Zhu, and X. Li, "MMGPL: Multimodal medical data analysis with graph prompt learning," *Med. Image Anal.*, vol. 97, Oct. 2024, Art. no. 103225, doi: 10.1016/j.media.2024.103225.
21. W. Liang, P. D. Meo, Y. Tang, and J. Zhu, "A survey of multi-modal knowledge graphs: Technologies and trends," *ACM Comput. Surv.*, vol. 56, no. 11, pp. 1–41, 2024, doi: 10.1145/3656579.
22. Y. Li et al., "KnowPhish: Large language models meet multimodal knowledge graphs for enhancing reference-based phishing detection," in *Proc. 33rd USENIX Secur. Symp.*, 2024, pp. 793–810.
23. S. Yang, R. Zhang, S. M. Erfani, and J. H. Lau, "UniMF: A unified framework to incorporate multimodal knowledge bases into end-to-end task-oriented dialogue systems," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 3978–3984.
24. M. D. Luecken et al., "A sandbox for prediction and integration of DNA, RNA, and proteins in single cells," in *Proc. 35th Conf. Neural Inf. Process. Syst. (NIPS)*, 2021.
25. J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "Genia corpus—A semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. suppl_1, pp. i180–i182, 2003, doi: 10.1093/bioinformatics/btg1023.
26. C. Nédellec et al., "Overview of bioNLP shared task 2013," in *Proc. BioNLP Shared Task Workshop*, 2013, pp. 1–7.

27. H. Ma et al., "Calibrating multimodal learning," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 23,429–23,450.
28. Y. Chen, J. Wang, L. Lin, Z. Qi, J. Ma, and Y. Shan, "Tagging before alignment: Integrating multi-modal tags for video-text retrieval," *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, 396–404, 2023, doi: 10.1609/aaai.v37i1.25113.
29. Q. Zhang et al., "Provable dynamic fusion for low-quality multimodal data," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 41,753–41,769.
30. B. Kim, S. Choi, D. Hwang, M. Lee, and H. Lee, "Transferring pre-trained multimodal representations with cross-modal similarity matching," in *Proc. 36th Conf. Neural Inf. Process. Syst. (NIPS)*, 2022, pp. 30826–30839.

CIYUAN PENG is a Ph.D. candidate at the Institute of Innovation, Science and Sustainability, Federation University Australia, Ballarat, VIC, 3350, Australia. Her research interests include brain science, digital health, and graph learning. Peng holds a master's degree in software engineering from Chung-Ang University, Seoul, South Korea. Contact her at ciyuan.p@ieee.org.

ESTRID HE is a senior lecturer at the School of Computing Technologies, RMIT University Melbourne, VIC, 3000, Australia. Her research interests include advancing artificial intelligence by improving deep learning models in terms of efficiency, scalability, and interpretability. He holds a Ph.D. degree in computer science from University of Melbourne, Australia. Contact her at estrid.he@rmit.edu.au.

FENG XIA is a professor in School of Computing Technologies, RMIT University, Australia, Melbourne, VIC, 3000, Australia. His research interests include artificial intelligence, graph learning, brain, robotics, and cyber-physical systems. Xia holds a Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China. He is a Fellow of the IEEE. He is the corresponding author of this article. Contact him at f.xia@ieee.org.