International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02066-y

Integrated Heterogeneous Graph Attention Network for Incomplete Multi-modal Clustering

Yu Wang[1,2,3] · Xinjie Yao[1,2,3] · Pengfei Zhu[1,2,3] · Weihao Li[4] · Meng Cao[1] · Qinghua Hu[1,2,3]

Received: 11 February 2023 / Accepted: 8 March 2024

Abstract

Incomplete multi-modal clustering (IMmC) is challenging due to the unexpected missing of some modalities in data. A key to this problem is to explore complementarity information among different samples with incomplete information of unpaired data. Despite preliminary progress, existing methods suffer from (1) relying heavily on paired data, and (2) difficulty in mining complementarity on data with high missing rates. To address the problems, we propose a novel method, Integrated Heterogeneous Graph ATtention (IHGAT) network, for IMmC. To fully exploit the complementarity among different samples and modalities, we first construct a set of integrated heterogeneous graphs based on the similarity graph learned from unified latent representations and the modality-specific availability graphs formed by the existing relations of different samples. Thereafter, the attention mechanism is applied to the constructed integrated heterogeneous graph to aggregate the embedded content of heterogeneous neighbors for each node. In this way, the representations of missing modalities can be learned based on the complementarity information of other samples and their other modalities. Finally, the consistency of probability distribution is embedded into the network for clustering. Consequently, the proposed method can form a complete latent space where incomplete information can be supplemented by other related samples via the learned intrinsic structure. Extensive experiments on eight public datasets show that the proposed IHGAT outperforms existing methods under various settings and is typically more robust in cases of high missing rates.

Keywords Incomplete multi-modal clustering · Integrated heterogeneous graph · Graph attention

1 Introduction

Various kinds of real-world data are usually represented with different modalities, such as perception data of intelligent unmanned systems and medical diagnosis data (Yang et al., 2019b; Chen et al., 2019; Cao et al., 2022). Among researches on modeling such multi-modal data, multi-modal clustering (MmC), which divides samples into clusters in an unsupervised manner, has attracted much attention in recent years (Zhang et al., 2020; Chen et al., 2022). MmC aims to integrate multiple features and discover complementary information among different modalities (Zhang et al., 2018; Xie et al., 2019; Fang et al., 2023). Compared with single-modality clustering, MmC can more fully exploit the complementarity between multiple modalities to improve performance (Han et al., 2023; Zhan et al., 2018). In real-world appli-

Communicated by Massimiliano Mancini.

✉ Pengfei Zhu
zhupengfei@tju.edu.cn

Yu Wang
wang.yu@tju.edu.cn

Xinjie Yao
yaoxinjie@tju.edu.cn

Weihao Li
whli@bu.edu

Meng Cao
caomeng@tju.edu.cn

Qinghua Hu
huqinghua@tju.edu.cn

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China

[2] Engineering Research Center of City Intelligence and Digital Governance, Ministry of Education of the People's Republic of China, Tianjin, China

[3] Haihe Lab of ITAI, Tianjin, China

[4] Department of Computer Science, Boston University, Boston, USA

Published online: 24 April 2024

cations, some modalities of instances may be missing due to the difficulty of data collection or the failure of data collectors (Kumar et al., 2013; Xiang et al., 2013). When certain modalities are missing, it leads to a significant loss of information. Furthermore, the absence of modalities severely hinders exploring complementary and consistent information. This indicates that incomplete multi-modal clustering presents its unique challenges (Wen et al., 2023; Lin et al., 2023). Such incompleteness further aggravates the difficulty of mining complementary information that can be originally mined through complete paired data. Therefore, how to effectively model the complementarity within incomplete data is an essential problem for incomplete multi-modal clustering (IMmC). The traditional MmC pipeline fails to address the challenges of IMmC. The core focus of research in the domain of missing modality multi-modal learning is to understand the impact of missing modality on modeling and representation. Unsupervised tasks, including clustering, prioritize the discovery of underlying data structures and relationships without relying on label information, making it a more challenging task. When some modalities are missing in the data, unsupervised tasks are generally more sensitive and capable of capturing these changes since they are not constrained by label information. In contrast, supervised tasks primarily focus on establishing a mapping between data and labels.

Many researchers have dedicated themselves to addressing the problem, and existing works can be roughly classified into three categories. (1) Grouping strategies divide data into multiple groups and design different models for each group. Then, these models are fused to alleviate the influence of missing modalities to obtain the clustering results (Yuan et al., 2012; Wang et al., 2020b). However, the amount of data used in this method for training is drastically reduced, which may lead to over-fitting. To alleviate the scarcity of complete modalities, the researchers proposed data imputation-based strategies. (2) Data imputation-based strategies complete the missing modalities of samples for the subsequent clustering, which transforms IMmC to a classical multi-modal clustering problem with complete data (Zhang et al., 2018; Lin et al., 2021). However, it is difficult to ensure the quality of the complete modality and may introduce additional noisy information, especially when the rate of missing data increases. To get rid of the reliance on large-scale complete data, recent studies have attempted to explore consistency in IMmC. (3) Consistency strategies (Zhang et al., 2022; Wang et al., 2021) generate missing modalities of samples by maintaining the consistent relationships between different modalities for the whole data. Although they reduce the requirement of paired data, the training process is quite unstable and is difficult to converge if the data distribution is complicated, *e.g.*, data with high missing rates. Consequently, the quality of the generated data is still difficult to control, which significantly deteriorates the performance of the models.

By revisiting existing methods, we find that two problems are still open: (1) *Modeling without relying heavily on paired data.* Grouping and data imputation-based strategies require a large number of paired data to learn the relationships between different modalities. For cases where only few complete data are available, these methods struggle to complete the missing modalities in high quality, thus deteriorating their performance. (2) *Mining complementarity on data with high missing rates.* Consistency strategies tend to learn relationships independently for each modality and can work well on simple cases, *e.g.*, data with low missing rates, with stable learning and convergence progress. However, the learned modality representations and structures become inaccurate when handling data with high missing rates, thus making them quite restricted to complicated real-world applications.

To this end, we propose a simple yet effective method, Integrated Heterogeneous Graph ATtention (IHGAT) network, to effectively and stably explore the structural information of samples and modalities without paired data. First, a set of integrated heterogeneous graphs is constructed by fusing two types of graphs: the similarity graph learned from unified latent representations and the modality-specific availability graphs obtained by the existing relations of different samples. Then, we adopt graph learning to exploit complementary structural information between samples based on the constructed integrated heterogeneous graphs. Concretely, we apply an attention mechanism to aggregate the embedded content of heterogeneous neighbors for each graph node. In this way, the incomplete data are embedded into a complete latent space while exploiting the structural information and maintaining the modality-missing information. Finally, the consistency of probability distribution is embedded into the network through KL divergence for clustering.

The proposed method has two advantages over existing methods, thus facilitating solving the aforementioned problems. (1) *Low dependency on the complete data.* The proposed method exploits the complementarity information by constructing a set of integrated heterogeneous graphs in a learnable unified feature space, where the relationships between different samples and modalities can be directly measured by their similarity. Such a simple method avoids the requirement for complete modalities of paired data. (2) *Effective exploitation of intrinsic structural information.* Based on the unified latent representation and constructed heterogeneous graphs, the proposed method aggregates the embeddings of heterogeneous neighbors for each node using an attention mechanism. In this way, the structural information and intra-sample and inter-sample multi-modal relationships can be fully exploited to enhance the capabilities of representation learning for samples with incomplete modalities.

Six common datasets with different missing rates are used in our experiments, and the results show that our method achieves state-of-the-art performance. Typically, our method is more robust than the baselines in cases of data with high missing rates, which infers that it can learn complementarity information between samples and modalities by learning intrinsic structural information without many paired data. The proposed method does not include any completion parts and is easy to implement. Source code is available at https://github.com/yxjdarren/IHGAT.

In summary, our contributions to this work are summarized as follows:

- We propose a structured information mining strategy, which involves constructing a heterogeneous graph structure within the data. This approach allows for the comprehensive exploration and exploitation of inter-modality and inter-sample relationships, facilitating the effective representation of incomplete data.
- The inter-modality relationships are realized by mapping multiple modalities into a unified latent space, while the inter-sample relationships are established based on the similarity within the latent space and the incomplete modality information. Such relationships are further used to facilitate complementary fusion among similar samples with graph attention mechanisms.
- Extensive experiments demonstrate the effectiveness of the proposed method on IMmC. Our method can maintain outstanding performance compared to the state-of-the-art baselines as the missing rate increases. Typically, IHGAT is significantly effective in scenarios with high missing rates, improving the baselines by up to 14.78 and 15.36% on Accuracy (ACC) and Normalized Mutual Information (NMI), respectively.

The remainder of this paper is organized as follows. In Sect. 2, we first review the related works about incomplete multi-modal learning and graph representation learning. Then, we elaborate on details of our work, including basic notations, framework, and analysis of each module in Sect. 3. Next, the experimental setting and evaluation results are reported in Sect. 4. Finally, we conclude our work in Sect. 5.

2 Related Work

In this section, we briefly review the related incomplete multi-modal learning and heterogeneous graph learning.

2.1 Incomplete Multi-Modal Learning

In contrast to multi-modal learning, incomplete multi-modal learning contains some missing modalities in data. Existing methods can be mainly categorized into three groups: grouping strategies, data imputation-based strategies, and consistency strategies.

Grouping strategies learn multiple models on various groups for late fusion and focus on the use of completeness theory (Baltrušaitis et al., 2018), which emphasizes complementarity to learn better latent representations. Specifically, within each group, samples with missing modalities are removed, resulting in multiple sets of complete multi-modal samples. However, the process of removing samples with missing modalities substantially reduces the available training data. Yuan et al. (2012) proposed to divide samples according to the availability of data sources and learn a base classifier for each data source independently. Xu et al. (2015) assumed that different modalities are generated from a shared subspace and investigated a successive over-relaxation method to solve the objective function. Wang et al. (2020b) proposed a framework based on knowledge distillation, utilizing the supplementary information from all modalities. However, the above methods have a relatively small amount of data in each group due to grouping, which may lead to overfitting. To address the lack of complete modalities, the data imputation-based strategies have attracted significant attention from researchers.

Data imputation-based strategies first complete the missing modalities, and then apply a common MmC algorithm. Enders (2010) simply imputed missing parts with the average value of all samples in each modality. Tran et al. (2017) imputed missing modalities by stacking residual autoencoders, which grows iteratively to model the residual between the current prediction and original data. Zhang et al. (2018) exploited the identical distribution constraint of missing modality to the other available one in the feature-isomorphic subspace to accomplish missing modality completion. Lin et al. (2021) proposed a novel objective that incorporates representation learning and data recovery into a unified framework from the modality of information theory. However, due to the noise that can be introduced when completing modalities, there is a trend to explore consistency between modalities.

Consistency strategies contain matrix factorization and consensus learning based IMmC to learn a consistent representation for different modalities. Hotelling (1992) proposed a matrix completion method by iterative soft thresholding of singular value decomposition. Shao et al. (2015) proposed Multi-Incomplete-modality Clustering (MIC), an algorithm based on weighted nonnegative matrix factorization with $L_{2,1}$ regularization. Zhang et al. (2022) proposed a novel framework to achieve the optimal tradeoff between consistency and complementarity across different modalities. Wang et al. (2021) proposed a generative partial multi-modal clustering model with adaptive fusion and cycle consistency, and a weighted adaptive fusion scheme was implemented to exploit the complementary information. Wang et al.

(2020) maximized the intrinsic correlations among different modalities by deep canonical correlation analysis to learn a consistent subspace representation among incomplete cross-modal data. While the above methods can explore inter-modality information with less complete data, they cannot stably learn and be convergent when dealing with data with high missing rates.

In real-world applications, massive paired data are hardly collected, and large portions of data may be missing due to the impact of environmental interference. In contrast to existing incomplete multi-modal learning methods, our proposed method requires less paired data and can handle cases of data with high missing rates, thus being capable of adapting to and working in such an open environment easily.

2.2 Graph Representation Learning

Graph learning is able to provide valuable insights into the structure of the data (Brasó et al., 2022; Brissman et al., 2023; Michieli & Zanuttigh, 2022). Li et al. (2021) jointly constructed local incomplete graph matrices, generated incomplete base partition matrices, stretched them to produce a unified partition matrix, and employed them to learn a consensus graph matrix. Wen et al. (2021) proposed a novel method introducing the tensor low-rank representation constraint and semantic consistency-based graph constraint. Cheng et al. (2020) designed Multi-View Attribute Graph Convolution Networks (MAGCN) with two-pathway encoders that map graph embedding features and learn modality-consistency information. Since MAGCN was designed assuming all modalities were fully and adequately observed, the design of its reconstruction loss functions and geometric consistency loss functions heavily relied on data completeness. Wen et al. (2020) developed a joint framework for graph completion and consensus representation learning, which introduces some adaptive weights to balance the importance of different modalities during consensus representation learning.

Unlike homogeneous graphs, attribute information is integrated into the clustering analysis on heterogeneous graphs. Heterogeneous graph learning is to learn effective representation from data of different attributes that are organized in multiple relation graphs (Wang et al., 2019; Zhang et al., 2019). Usually, constructing heterogeneous graphs requires considering the difference in neighbor information under different relationships. Therefore, heterogeneous graph neural networks usually adopt hierarchical aggregation. To implement the hierarchical aggregation function, heterogeneous graphs usually need to consider the difference in neighbor information under different relationships (Chang et al., 2015; Zhang et al., 2018c).

Different from traditional homogeneous graph structure learning, considering the heterogeneity of different relations in the heterogeneous graph, heterogeneous graph structure learning (Zhao et al., 2021) generates each relation subgraph separately. At present, there are relatively few studies on heterogeneous graph learning applied to IMmC (Bothorel et al., 2015; Shi et al., 2016). Qi et al. (2012) proposed heterogeneous random fields to model the structure and content of social media networks. Li et al. (2017) studied the problem of clustering objects in an attributed heterogeneous information network, taking into account the similarities of objects with respect to both object attribute values and their structural connectedness in the network. Chen et al. (2020) represented attributed graphs as star-schema heterogeneous graphs to capture both structural and attribute similarities, where attributes are modeled as different types of graph nodes. Yang et al. (2019a) learned the common subspace with the adaptive graph fusion, which allows the integration of complementary and consistent information from different modalities.

In our work, heterogeneous graphs are constructed to mine the complementarity information between samples and modalities deeply. Unlike some usual graph representation learning methods, we consider the heterogeneity of relations of different modalities and samples by fusing the similarity graph and modality-specific availability graphs. By learning representations based on the heterogeneous graphs, the structural information inside the incomplete multi-modal data is learned to exploit the complementarity between different samples and modalities, which yields compact representations for incomplete multi-modal clustering.

3 Methodology

In this section, the details of the proposed method are illustrated. We design an integrated heterogeneous graph attention network that includes a latent representation learning layer, an integrated heterogeneous graph construction layer, and a clustering layer (See Fig. 1). First, a set of integrated heterogeneous graphs is constructed by fusing: (1) the similarity graph that reflects the neighborhood relations of samples, and (2) the modality-specific availability graphs that encode the modality-existence information. Then, the attention mechanism is applied to the obtained graphs to learn complete representations of data. Finally, considering the consistency of the probability distribution, we use KL divergence to measure the non-symmetric difference between two probability distributions and obtain the clustering results. Details of different modules are presented in the following subsections.

Problem Definition. Consider data $\{\mathbf{S}_n\}_{n=1}^{N}$, where $\mathbf{S}_n$ is a subset of the complete observations $\mathbf{X}_n = \{x_n^{(v)}\}_{v=1}^{V}$ (*i.e.*, $\mathbf{S} \subset \mathbf{X}$) with N and V being the number of samples

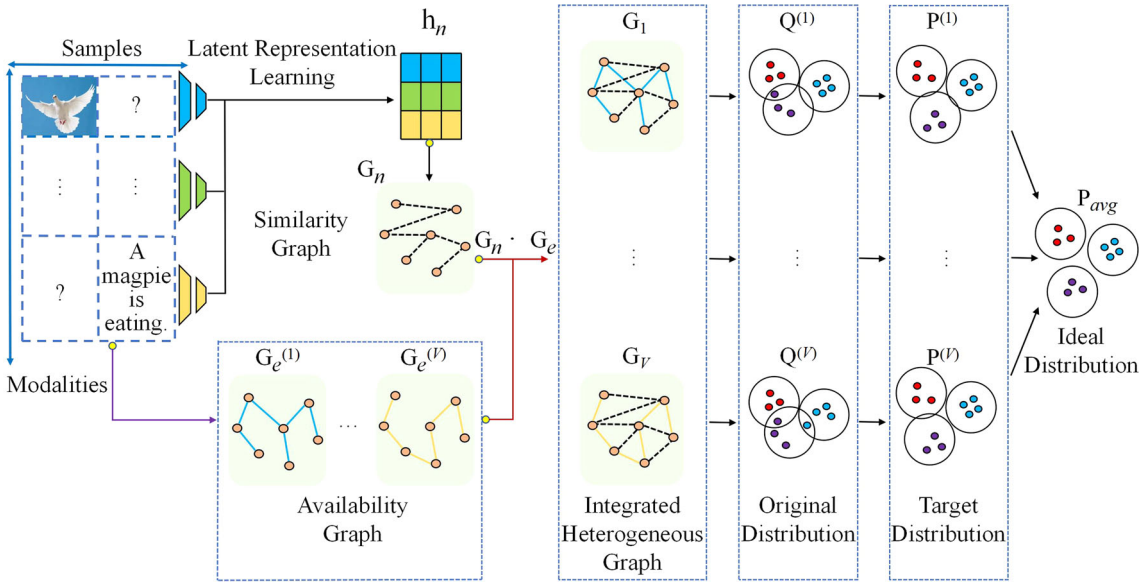

Fig. 1 The architecture of our method. IHGAT requires only a small amount of paired data to model missing modalities, focusing more on internal structural information rather than using modality-completion methods that may introduce noise. Firstly, we construct a set of integrated heterogeneous graphs based on the similarity graph learned from unified latent representations and the modality-specific availability graphs obtained by the existing relations of different samples. Next, we apply an attention mechanism to aggregate the embedded content of heterogeneous neighbors for each node. Finally, the consistency of probability distribution is embedded into the network for clustering

and modalities, respectively. $\mathbf{X}_n \in \mathbf{R}^{N \times V}$ is the n-th sample with all modalities. IMmC aims to cluster data in which some samples have missing modalities so that samples $\mathbf{S}_n$ with the arbitrary possible missing-modalities pattern can be clustered.

3.1 Integrated Heterogeneous Graph Construction

The integrated heterogeneous graphs are composed of the similarity graph and the modality-specific availability graphs. We use the consistency loss $\mathcal{L}_c$ to measure the non-symmetric difference between the original distribution and the target distribution. By embedding $\mathcal{L}_c$ into the network, we can simultaneously optimize both reconstruction loss $\mathcal{L}_r$ and consistency loss $\mathcal{L}_c$ within a unified framework. This approach offers the advantage of allowing the network to capture the intrinsic structure of data better while capturing the complementarity between samples and modalities through integrated heterogeneous graph attention networks, thereby improving the performance of the model. Next, we elaborate on the construction of each graph.

3.1.1 Similarity Graph

Similar samples can help each other for representation learning, and they should be close in the learned latent space. To this end, we construct the similarity graph to maintain the local structure of the data by first learning unified latent representations of all modalities and then obtaining the graph based on the similarity of samples in the latent space.

To process samples with arbitrary missing-modality modes flexibly, we project the samples into a unified latent space. Ideally, the expression of the hidden layer can extract the unified expression from each modality. If we denote the latent space representation of the n-th sample as $\mathbf{h}_n$, then the optimization objective of the elastic implicit space representations is as follows:

$$\mathcal{L}_r(s_{nv}, \mathbf{S}_n, \mathbf{h}_n; \Theta_r) = \sum_{n=1}^{N} \sum_{v=1}^{V} s_{nv} \left\| f_v\left(\boldsymbol{h}_n; \boldsymbol{\Theta}_r^{(v)}\right) - \boldsymbol{s}_n^{(v)} \right\|^2, \tag{1}$$

where $\mathcal{L}_r$ is the reconstruction loss, which aims to learn the bidirectional mapping between the original data space and the unified embedding space. $\|\cdot\|$ represents the l_2-norm. $f_v\left(\mathbf{h}_n; \Theta_r^{(v)}\right)$ is the reconstruction network for the v-th modality parameterized by $\Theta_r^{(v)}$, and $\mathbf{s}_n^{(v)}$ represents the input of the v-th modality with the n-th sample. N and V represent the number of samples and modalities, respectively. $\mathbf{s}_{nv}$ indicates the availability of the n-th sample in the v-th modality, which is defined as follows:

$$\mathbf{s}_{nv} = \begin{cases} 1, & \text{if the } n\text{-th instance has the } v\text{-th modality} \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

3.1.2 Learnable Unified Latent Representation

By using multiple individual multi-layer perceptrons as encoders, different available modalities are encoded into a unified learnable space $\mathbf{h}_n$ (regardless of their lost patterns), where the number of encoders should be the same as modalities. Relatively complete and universal representations are learned by minimizing Eq. (1), so that any sample with missing patterns can be reconstructed. This means that the space has learned the potential elastic representations from the observation modality.

Generally, the neighborhood structure can be obtained from a Gaussian-based kernel matrix. We denote the matrix as $\mathbf{G}_n \in \mathbf{R}^{N\times N}$, and the detailed formulation is as follows:

$$\mathbf{G}_{\mathrm{n}ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{h}_i-\mathbf{h}_j\|^2}{2\sigma^2}\right), \mathbf{h}_i \in N_k\left(\mathbf{h}_j\right) \text{ or } \mathbf{h}_j \in N_k\left(\mathbf{h}_i\right) \\ 0, \text{ otherwise} \end{cases}, \tag{3}$$

where σ is the standard deviation. $N_k\left(\mathbf{h}_i\right)$ and $N_k\left(\mathbf{h}_j\right)$ indicate samples of the K nearest neighbors of $\mathbf{h}_i$ and $\mathbf{h}_j$, respectively.

3.1.3 Modality-Specific Availability Graphs

For different modalities, the absence of internal samples may vary. Two different samples in the same modality can only interact with each other if they exist at the same time. We propose modality-specific availability graphs based on multiple samples in the same modality to make full use of the similarities between samples. We denote the matrices as $\mathbf{G}_e \in \mathbf{R}^{V\times N\times N}$ and $\mathbf{G}_e^{(v)} \in \mathbf{R}^{N\times N}$. The detailed formulation is defined as follows:

$$\mathbf{G}_{e_{ij}}^{(v)} = \begin{cases} 1, \text{ if both } x_i^{(v)} \text{ and } x_j^{(v)} \text{ exist} \\ 0, \text{ otherwise} \end{cases}, \tag{4}$$

where $x_i^{(v)}$ and $x_j^{(v)}$ are the v-th modality of different samples,

$$\mathbf{G}_e = \left[\mathbf{G}_e^{(1)}, \mathbf{G}_e^{(2)}, \mathbf{G}_e^{(3)}, ..., \mathbf{G}_e^{(V)}\right]. \tag{5}$$

3.1.4 Integrated Heterogeneous Graph

To further utilize the complementarity of data information, we consider fusing the similarity graph and the modality-specific availability graphs to obtain a set of integrated heterogeneous graphs. We define the graph adjacency matrix as follows:

$$\mathbf{G}_{adj} = \mathbf{G}_n \cdot \mathbf{G}_e^{(v)}. \tag{6}$$

Then, we obtain:

$$\mathbf{G} = \left[\mathbf{G}_n \cdot \mathbf{G}_e^{(1)}, \mathbf{G}_n \cdot \mathbf{G}_e^{(2)}, ..., \mathbf{G}_n \cdot \mathbf{G}_e^{(V)}\right]. \tag{7}$$

3.2 Graph Representation Learning

Given the integrated heterogeneous graphs, we can exploit structural information to learn complete representation. Our research is focused on addressing the challenges of incomplete multi-modal learning, with a particular emphasis on harnessing the interrelationships between modalities and samples in the absence of partial modalities. In other words, our intention is not to introduce a novel attention mechanism but rather to make the most of existing methodologies to explore the interrelationships in data with incomplete modalities in a comprehensive manner. Leveraging the dynamic adaptability of the attention mechanism introduced by Graph Attention Network (GAT) (Veličković et al., 2018), we have seamlessly incorporated GAT into our framework for the purpose of exploring the interrelationships between modalities and samples. By leveraging masked self-attention layers and stacking layers, nodes can attend to the features of their neighborhoods.

Formally, given the latent representations $\mathbf{h}_n = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\}$, where $\mathbf{h}_n \in \mathbf{R}^D$, N is the total number of samples, and D is the dimension of latent space, the network outputs the adjacency matrix $\mathbf{G}_v$ generated by multiple graph learners based on different samples. Each of these samples is then possible to generate V groups of new features $\mathbf{z}^{(v)} = \left\{z_1^{(v)}, z_2^{(v)}, \ldots, z_N^{(v)}\right\}\left(z_n^{(v)} \in \mathbf{R}^F, n = 1, 2, \ldots, N\right)$, where N is the number of nodes, F is the number of features in each node and $z_n^{(v)}$ refers to the feature vector associated with the n-th node and the v-th modality.

To facilitate graph representation learning, we use GAT to transform the latent representations into features that are suitable for graph semantics, which requires a mapping layer composed of learnable parameters $\Theta_g^{(v)}$:

$$z_n^{(v)} = GAT\left(\mathbf{h}_n, \mathbf{G}_v; \Theta_g^{(v)}\right), \tag{8}$$

where $\mathbf{h}_n$ is the latent representations, $\mathbf{G}_v$ is the graph adjacency matrix, and $\Theta_g^{(v)}$ is the parameter set of the GAT.

Based on representations $\mathbf{z}^{(v)}$, we use the attention mechanism to calculate the importance between nodes. As an initial step, a shared linear transformation, parametrized by a weight matrix, $\mathbf{W} \in \mathbf{R}^{F'\times F}$ (of potentially different cardinality F'), is applied to every node. The importance of each two nodes can be calculated by a shared attention mechanism a. Thus, attention coefficients are defined as:

$$e_{ij} = a\left(\mathbf{W}z_i^{(v)}, \mathbf{W}z_j^{(v)}\right). \tag{9}$$

The attention mechanism a is a feedforward network, parametrized by a weight vector $\overrightarrow{\mathbf{a}} \in \mathbf{R}^{2F'}$, and e_{ij} indicates the importance of node j to node i. We use an attention mask to inject the integrated heterogeneous graph structure into the calculation, and only consider e_{ij} for nodes $j \in \mathcal{N}_i$ that have relations in $\mathbf{G}_v$. Projected by a softmax function, the formula is:

$$\alpha_{ij} = \text{softmax}_j\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(e_{ik}\right)}, \tag{10}$$

where $\mathcal{N}_i$ is some neighborhood of node i in the graph. The attention mechanism uses a single-layer feedforward network, and applies the LeakyReLU nonlinearity. Fully expanded out, the attention calculation can be expressed as:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\overrightarrow{\boldsymbol{a}}^{\mathrm{T}}\left[\boldsymbol{W}\mathrm{z}_{\mathrm{i}}^{(\mathrm{v})}\boldsymbol{W}\mathrm{z}_{\mathrm{j}}^{(\mathrm{v})}\right]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\overrightarrow{\boldsymbol{a}}^{\mathrm{T}}\left[\boldsymbol{W}\mathrm{z}_{\mathrm{i}}^{(\mathrm{v})}\boldsymbol{W}\mathrm{z}_{\mathrm{k}}^{(\mathrm{v})}\right]\right)\right)}, \tag{11}$$

where $\cdot^T$ represents transposition and $[\cdot\|\cdot]$ is the concatenation operation. Moreover, multi-head attention can be used to enrich the ability of the method and stabilize the training process. Each head of attention has its own parameters. We use splicing to integrate the output of multiple attention mechanisms, which can be described as follows:

$$z_i^{\prime(v)} = \|_{b=1}^{B}\mu\left(\sum_{j\in N_i}\alpha_{ij}^{b}\mathbf{W}^{b}z_j^{(v)}\right), \tag{12}$$

where $\|$ represents concatenation, μ is the activation function, α_{ij}^{b} are normalized attention coefficients computed by the b-th attention mechanism (a^b), B is the number of attention heads and $\mathbf{W}^b$ is the corresponding weight matrix of input linear transformation. When dealing with the last layer, we use the average value instead of the concatenation, as follows

$$z_i^{\prime(v)} = \mu\left(\frac{1}{B}\sum_{b=1}^{B}\sum_{j\in N_i}\alpha_{ij}^{b}\mathbf{W}^{b}z_j^{(v)}\right), \tag{13}$$

3.2.1 Consistent Embedding Network

Based on the learned features, we optimize the clustering task in an end-to-end manner. To measure the non-symmetric difference between the original distribution and target distribution, we embed the consistency of probability distribution into the network. Following (Wang et al., 2018; Tao et al., 2019), we measure the similarity between integrated node representations $z_i^{\prime(v)}$ and the cluster center μ_j by adopting the t-distribution of student. $q_{ij}^{(v)}$ and $p_{ij}^{(v)}$ are the elements of original distribution $\mathbf{Q}$ and target distribution $\mathbf{P}$, respectively, which is defined as:

$$q_{ij}^{(v)} = \frac{(1+\|z_i^{\prime(v)}-\mu_j\|^2/\beta)^{-\frac{\beta+1}{2}}}{\sum_{j'=1}^{J}(1+\|z_i^{\prime(v)}-\mu_{j'}\|^2/\beta)^{-\frac{\beta+1}{2}}}, \tag{14}$$

where $\|\cdot\|$ represents the l_2-norm; μ_j is the cluster center; J is the number of cluster centers; β is the degree of t-distribution freedom of the Student, and $q_{ij}^{(v)}$ is the probability of assigning node i to cluster j. In our experiments, the cluster centers $\{\mu_j\}_{j=1}^{J}$ can be initialized by employing k-means and the target probability distribution $p_{ij}^{(v)}$ ($0 \le p_{ij}^{(v)} \le 1$) can be computed. We obtain $p_{ij}^{(v)}$ by raising $q_{ij}^{(v)}$ to the second power and normalizing by frequency per cluster:

$$p_{ij}^{(v)} = \frac{{q_{ij}^{(v)}}^2/f_i}{\sum_{j'=1}^{J}{q_{ij'}^{(v)}}^2/f_{j'}}, \tag{15}$$

where $f_j = \sum_{i=1}^{M} q_{ij}^{(v)}$ are soft cluster frequencies. To compare the similarity of the two probability distributions, we define our objective as a probability distribution consistency loss $\mathcal{L}_c$. The clustering loss is defined as minimizing the KL divergence between an original distribution and a target distribution. That is to say, $\mathcal{L}_c$ is defined as:

$$\mathcal{L}_c(\mathbf{P},\mathbf{Q}) = KL(\mathbf{P}\|\mathbf{Q}) = \sum_{v}\sum_{i}\sum_{j}p_{ij}^{(v)}log\frac{p_{ij}^{(v)}}{q_{ij}^{(v)}}, \tag{16}$$

where KL is the Kullback–Leibler divergence that measures the non-symmetric difference between two probability distributions. $(\cdot\|\cdot)$ represents the separator between two probability distributions. $\mathbf{P}$ and $\mathbf{Q}$ are defined by Eq. (15) and Eq. (14), respectively. Finally, we take the mean of $\{p_{ij}^{(v)}\}_{v=1}^{V}$ as the ideal distribution.

Accordingly, the overall loss function of the proposed IHGAT can be formulated as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}_c + \lambda_2\mathcal{L}_r, \tag{17}$$

where $\mathcal{L}_c$ and $\mathcal{L}_r$ are the clustering loss and the reconstruction loss, respectively. The λ_1 and λ_2 are the trade-off hyperparameters of the $\mathcal{L}_c$ and $\mathcal{L}_r$, respectively.

We construct a set of integrated heterogeneous graphs based on the similarity graph learned from unified latent representations and the modality-specific availability graphs obtained by the existing relations of different samples. Based

Algorithm 1 Algorithm for IHGAT

Input: Incomplete multi-modality dataset $\{\mathbf{S}_n\}$, the dimensionality of latent representations γ
and the number of nearest neighbor samples K.

Output: the parameters of the model $\left\{\Theta_r^{(v)}\right\}_{v=1}^{V}$, $\left\{\Theta_g^{(v)}\right\}_{v=1}^{V}$, and the learned representations of samples $\left\{\mathbf{z}'^{(v)}\right\}_{v=1}^{V}$.

for $v = 1$ to V **do**
  $\Theta_r^{(v)} \leftarrow \Theta_r^{(v)}$-$\partial L/\partial \Theta_r^{(v)}$
end for
for $n = 1$ to N **do**
  Update $\mathbf{h}_n$ with gradient descent:
  $\mathbf{h}_n \leftarrow \mathbf{h}_n$-$\partial L/\partial \mathbf{h}_n$
end for
for $v = 1$ to V **do**
  Update $\Theta_g^{(v)}$ with gradient descent:
  $\Theta_{\mathbf{g}}{}^{(v)} \leftarrow \Theta_{\mathbf{g}}{}^{(v)}$-$\partial L/\partial \Theta_{\mathbf{g}}{}^{(v)}$
end for

on the constructed integrated heterogeneous graphs, we use the incomplete multi-modal data as input to optimize the model parameters for a better representation. Algorithm 1 briefly summarizes the optimization procedures of the proposed method.

4 Experiments

In this section, we conducted comprehensive experiments on incomplete multi-modal data to evaluate the performance of our proposed method, followed by the analysis of our proposed method.

4.1 Metrics and Datasets

For a comprehensive analysis, we conducted extensive experiments on six datasets and adopted two widely used metrics, including Accuracy (ACC) and Normalized Mutual Information (NMI). High values denote good clustering performance of the method for both metrics.

CUB (Wah et al., 2011): Caltech-UCSD Birds (CUB) contains 11,788 bird images associated with text descriptions from 200 different categories (we followed the experimental settings in Zhang et al. (2022), so the first 10 categories are used). We extracted 1024-dimensional features based on images using GoogLeNet, and 300-dimensional features based on text (Le & Mikolov, 2014).

Football[1]: A collection of 248 English Premier League football players and clubs active on Twitter. The disjoint ground-truth communities correspond to the 20 individual clubs in the league.

ORL[2]: ORL is a popular face database in the field of face recognition. It contains 400 face images provided by 40 volunteers, with 10 face images from each person. Three types of features, *i.e.*, LBP, Gabor, and intensity, are extracted as the three modalities for representing every face image.

PIE[3]: PIE is a subset containing 680 facial images of 68 subjects, for which the intensity, LBP, and Gabor features have been extracted.

Politics[4]: A collection of Irish politicians and political organizations, assigned to seven disjoint ground truth groups, according to their affiliation.

3Sources[5]: 3Sources is collected from three online news sources: BBC, Reuters, and Guardian. In total, 169 samples of stories are used, which are reported by all three sources.

ADNI[6]: The dataset consists of 774 subjects from ADNI-1, including 226 normal controls (NC), 362 MCI and 186 AD subjects. There are only 379 subjects with complete MRI and PET data, including 101 NC, 185 MCI, and 93-AD, where the missing rate is up to 0.26. We use 93-dimensional ROI-based features from both MRI and PET data, respectively.

3Sources-partial[7]: 948 news articles were collected covering 416 distinct news stories. Specifically, 169 were reported in all three sources, 194 in two sources, and 53 appeared in a single news source. Each source represents a unique modality, and the combination of different sources constitutes multi-modal information. The missing rate of 3Sources-partial is 0.24.

4.2 Experimental Setups

To generate incomplete multi-modal datasets from complete multi-modal datasets, we randomly removed different modalities within each sample based on the missing rate. The missing rate was defined as $\varepsilon = \frac{\sum_v M_v}{V \times N}$, where M_v indicates the number of instances without the v-th modality. To evaluate the influence of λ_1 and λ_2, we changed their value in the range of $\{0.01, 0.1, 1, 10, 100, 1000\}$ and $\{0.01, 0.1, 1, 10, 100, 1000\}$, respectively. As shown in Fig. 6, IHGAT was robust to λ_1 and λ_2, and the proposed method could reach a high level of performance while λ_1 was from the range of $\{10, 100, 1000\}$ and λ_2 was from the range of $\{0.01, 0.1, 1\}$. For all datasets, the trade-off hyper-parameters λ_1 and λ_2 were fixed to 100 and 1, respectively. We evaluated the performance and reported the averaged results over five runs of

[1] http://mlg.ucd.ie/aggregation/index.html.

[2] https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[3] http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html.

[4] http://mlg.ucd.ie/aggregation/index.html.

[5] http://mlg.ucd.ie/datasets/3sources.html.

[6] http://www.loni.usc.edu/ADNI.

[7] http://erdos.ucd.ie/datasets/3sources.html.

Table 1 The clustering performance comparison on six datasets with different missing rates (ε)

Dataset	Method	ACC (%)					NMI (%)				
		ε=0.1	ε=0.2	ε=0.3	ε=0.4	ε=0.5	ε=0.1	ε=0.2	ε=0.3	ε=0.4	ε=0.5
CUB	SVD [1992] (Hotelling, 1992)	71.43±1.91	60.46±1.83	51.87±1.23	46.48±1.43	40.96±1.15	67.88±0.89	58.05±0.02	50.29±0.92	48.44±0.61	43.87±0.53
	Average [2010] (Enders, 2010)	62.17±1.88	55.10±1.24	48.45±0.43	48.35±0.72	37.60±1.50	58.49±0.05	55.60±0.24	53.97±0.21	52.98±0.09	48.09±0.30
	CRA [CVPR 2017] (Tran et al., 2017)	70.19±1.86	59.83±1.79	52.41±1.56	51.41±1.72	40.30±0.08	67.51±0.85	58.31±0.77	53.11±1.55	52.89±1.94	45.39±0.24
	iCmSC [TIP 2020] (Wang et al., 2020)	68.02±2.28	66.17±1.69	60.42±1.84	52.83±1.93	41.92±2.01	63.28±1.63	62.19±1.82	57.38±2.02	54.17±1.83	45.21±2.06
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	*76.96±1.57*	69.67±1.87	63.32±1.97	55.75±2.65	45.25±1.83	*78.34±2.52*	69.22±1.75	62.31±1.85	59.38±2.49	52.46±1.87
	GP-MVC [TIP 2021] (Wang et al., 2021)	70.28±1.95	67.28±1.86	62.81±1.99	53.48±2.08	43.81±1.83	67.93±1.73	63.85±1.91	60.01±2.24	55.86±1.89	48.71±1.97
	CPM [TPAMI 2022] (Zhang et al., 2022)	71.76±1.60	68.15±2.48	61.13±2.98	53.11±2.31	42.33±1.66	72.82±0.72	67.72±1.61	60.14±1.42	55.49±0.91	46.92±1.48
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	74.61±2.55	*71.20±2.26*	*63.87±1.47*	54.67±2.76	45.74±0.17	71.98±0.89	69.31±0.86	62.69±0.70	56.58±1.43	49.74±0.68
	PIMVC[TNNLS 2023] (Deng et al., 2023)	74.08±3.16	69.72±2.81	61.36±2.92	55.98±2.75	51.99±2.56	74.72±2.86	68.85±2.69	63.05±2.88	58.97±2.91	57.86±3.02
	GreatF [TCSVT 2023] (Wen et al., 2023a)	75.13±2.91	70.01±2.48	62.24±2.87	*56.47±2.83*	*52.19±2.85*	76.08±2.72	*70.57±3.01*	*64.18±2.76*	*59.73±2.79*	*58.37±2.81*
	IHGAT	**78.01±1.83**	**72.52±2.18**	**65.91±1.64**	**59.33±2.00**	**55.42±2.32**	**78.36±2.22**	**73.35±2.13**	**66.54±1.91**	**61.99±2.46**	**60.82±2.13**
Football	SVD [1992] (Hotelling, 1992)	33.17±0.73	27.93±0.71	24.70±0.63	18.57±0.48	13.62±0.67	48.52±0.51	44.77±0.73	40.25±0.49	36.61±0.26	30.47±0.19
	Average [2010] (Enders, 2010)	36.39±1.79	29.36±1.38	26.45±1.13	22.00±0.87	18.59±1.32	48.17±0.68	40.35±0.35	36.44±0.82	30.27±1.02	25.54±0.28
	CRA [CVPR 2017] (Tran et al., 2017)	36.55±3.50	32.81±2.57	30.59±1.96	25.43±1.09	23.22±1.45	52.29±0.30	45.59±0.39	41.37±0.44	34.12±0.23	29.40±0.78
	iCmSC [TIP 2020] (Wang et al., 2020)	40.28±1.76	39.09±1.82	37.75±1.97	34.69±2.15	33.02±2.37	52.12±1.73	50.89±2.01	47.92±2.05	45.31±2.03	40.19±2.19
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	**73.05±1.65**	66.83±2.21	61.37±1.81	54.84±2.41	47.87±2.61	78.35±1.85	72.09±1.93	67.72±2.22	59.62±1.72	50.04±1.66
	GP-MVC [TIP 2021] (Wang et al., 2021)	42.05±1.59	40.92±1.86	39.95±1.81	36.89±2.02	35.71±2.29	53.95±1.85	52.18±1.78	50.16±2.21	48.33±2.19	41.97±1.99
	CPM [TPAMI 2022] (Zhang et al., 2022)	43.31±2.25	40.91±1.34	40.53±1.02	38.55±0.43	34.55±2.27	53.30±0.65	53.20±1.27	52.76±2.07	51.29±0.17	42.66±2.00
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	42.77±0.95	42.55±1.66	41.64±1.51	39.13±1.06	36.83±2.86	54.06±0.56	54.67±1.25	53.14±1.19	52.44±1.03	44.56±3.32
	PIMVC[TNNLS 2023] (Deng et al., 2023)	70.28±2.31	67.29±2.82	62.03±2.48	57.38±2.87	*53.99±2.93*	79.58±2.39	76.85±2.67	72.29±2.58	70.99±2.92	*67.93±2.81*
	GreatF [TCSVT 2023] (Wen et al., 2023a)	70.83±4.05	*68.98±3.83*	*62.18±2.95*	*57.89±2.67*	53.75±3.24	*80.01±3.97*	*77.83±2.54*	*73.18±2.93*	*71.82±2.89*	67.21±2.77
	IHGAT	*71.36±3.93*	**70.29±3.55**	**64.09±2.48**	**60.36±1.92**	**55.89±2.88**	**82.80±3.63**	**80.95±1.98**	**75.04±2.56**	**73.97±2.74**	**69.97±2.15**
ORL	SVD [1992] (Hotelling, 1992)	59.25±2.04	54.78±0.94	47.17±0.87	41.86±1.89	34.25±0.74	75.86±0.59	70.94±0.38	65.33±0.16	58.49±0.94	53.69±0.44
	Average [2010] (Enders, 2010)	55.25±1.77	47.61±1.18	40.64±1.47	36.83±1.54	31.97±0.66	72.36±0.58	65.81±0.42	59.36±0.61	56.03±0.32	52.09±0.65
	CRA [CVPR 2017] (Tran et al., 2017)	58.83±2.38	50.68±1.28	47.28±0.40	39.48±0.48	35.15±0.38	74.60±0.72	68.09±0.21	62.71±0.51	56.94±0.16	52.27±0.79
	iCmSC [TIP 2020] (Wang et al., 2020)	59.03±2.13	51.49±2.36	45.28±2.75	41.09±2.46	35.24±2.85	69.81±2.29	68.23±1.89	64.95±2.51	60.27±2.63	56.86±2.75
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	82.75±2.36	81.37±2.75	79.53±2.25	77.58±1.86	69.97±1.72	92.53±2.19	91.08±2.47	90.85±2.21	87.34±2.36	84.71±2.64
	GP-MVC [TIP 2021] (Wang et al., 2021)	60.28±1.92	53.01±2.05	47.22±2.13	42.31±2.42	36.98±2.55	72.07±1.92	69.57±2.13	66.08±2.36	61.97±2.45	57.99±2.35
	CPM [TPAMI 2022] (Zhang et al., 2022)	57.83±2.42	50.89±2.00	44.03±1.28	41.89±1.37	34.64±0.31	74.78±1.54	73.16±0.82	66.70±0.83	63.10±0.42	56.99±0.44
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	62.61±2.27	56.28±2.18	49.64±0.99	43.75±1.34	37.39±1.74	75.83±0.94	74.29±0.91	69.31±0.73	63.81±0.62	58.32±1.11
	PIMVC[TNNLS 2023] (Deng et al., 2023)	82.98±2.83	81.83±2.94	80.49±2.67	78.28±2.91	70.58±2.83	92.96±3.02	91.89±2.89	91.29±2.58	87.98±2.96	85.23±2.85
	GreatF [TCSVT 2023] (Wen et al., 2023a)	*83.09±2.89*	*82.78±2.96*	*81.02±2.99*	*78.73±2.82*	*71.29±3.02*	*93.18±2.83*	*91.99±2.84*	*92.17±2.83*	*91.02±2.93*	*87.89±2.93*
	IHGAT	**85.63±1.93**	**85.01±2.57**	**83.09±2.87**	**81.04±2.18**	**73.61±2.19**	**95.02±2.14**	**94.94±2.24**	**94.03±2.48**	**93.17±2.10**	**90.02±2.71**

Table 1 continued

Dataset	Method	ACC (%)					NMI (%)				
		ε=0.1	ε=0.2	ε=0.3	ε=0.4	ε=0.5	ε=0.1	ε=0.2	ε=0.3	ε=0.4	ε=0.5
PIE	SVD [1992] (Hotelling, 1992)	26.11±0.92	23.84±0.80	22.61±0.41	21.80±0.68	19.31±0.48	55.89±0.90	52.13±0.46	49.82±0.78	48.42±0.57	44.80±1.23
	Average [2010] (Enders, 2010)	26.16±0.77	24.38±0.43	23.15±0.62	21.88±0.82	19.15±0.23	56.03±0.61	51.71±0.62	49.28±0.57	47.97±1.04	43.52±0.90
	CRA [CVPR 2017] (Tran et al., 2017)	25.16±1.10	22.72±0.58	21.96±0.62	19.21±0.56	19.19±0.52	55.45±0.71	52.61±0.57	49.01±0.80	46.25±0.83	44.70±0.93
	iCmSC [TIP 2020] (Wang et al., 2020)	45.81±2.01	40.27±2.48	38.39±2.19	32.07±2.26	26.24±2.77	68.18±1.93	60.29±2.13	60.06±2.46	54.93±2.28	48.58±2.31
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	62.94±1.93	55.09±2.21	52.07±2.03	45.45±2.25	38.71±2.29	80.31±1.69	77.91±2.09	75.57±1.92	73.43±1.71	70.76±1.95
	GP-MVC [TIP 2021] (Wang et al., 2021)	46.03±1.53	41.56±1.66	39.18±1.82	32.98±1.79	26.89±2.16	70.02±1.61	62.18±1.82	60.87±1.86	54.98±1.96	49.82±2.14
	CPM [TPAMI 2022] (Zhang et al., 2022)	53.14±1.12	37.91±0.92	37.55±0.98	31.70±1.28	23.68±0.68	72.33±1.17	60.02±0.73	60.27±1.09	54.68±0.84	47.83±0.46
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	52.29±1.20	45.59±1.29	41.37±1.34	34.12±1.13	29.40±1.68	72.44±0.55	65.43±1.12	63.69±1.50	57.71±1.20	52.76±1.43
	PIMVC[TNNLS 2023] (Deng et al., 2023)	63.05±3.56	55.02±3.18	52.17±3.29	44.95±2.89	39.95±2.83	81.96±2.58	79.17±2.53	77.09±2.82	74.35±2.76	73.04±2.69
	GreatF [TCSVT 2023] (Wen et al., 2023a)	*63.81±3.81*	*55.89±3.92*	*52.96±3.87*	*45.27±2.93*	*40.01±2.75*	*82.19±2.89*	*79.88±2.97*	*77.53±2.78*	*74.78±2.57*	*73.51±2.42*
	IHGAT	**64.29±3.76**	**57.82±3.35**	**54.32±3.54**	**47.99±3.29**	**42.04±2.62**	**85.13±1.87**	**82.75±1.75**	**80.91±2.57**	**78.53±1.78**	**76.36±2.42**
Politics	SVD [1992] (Hotelling, 1992)	49.06±0.56	46.91±1.20	45.65±0.82	38.75±0.99	36.70±0.64	42.06±0.56	39.91±1.20	38.65±0.82	31.75±0.99	24.48±0.57
	Average [2010] (Enders, 2010)	49.64±1.97	44.75±1.58	42.45±1.75	39.93±0.95	38.26±1.83	45.47±1.03	40.12±1.38	34.93±1.87	26.30±1.60	22.04±1.90
	CRA [CVPR 2017] (Tran et al., 2017)	55.07±2.38	50.32±1.75	46.41±0.74	40.12±0.92	36.97±0.46	47.51±1.65	38.31±1.57	37.11±2.35	32.89±2.74	25.39±1.04
	iCmSC [TIP 2020] (Wang et al., 2020)	57.05±1.72	55.01±1.85	54.83±1.94	44.17±2.27	41.58±2.35	51.85±1.96	50.72±1.93	49.31±2.09	43.26±2.26	29.18±2.42
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	64.94±1.97	60.41±1.76	59.39±2.27	57.19±2.38	55.02±2.68	**68.62±2.52**	60.76±2.79	58.81±2.45	55.91±2.32	46.03±2.28
	GP-MVC [TIP 2021] (Wang et al., 2021)	58.96±1.81	57.93±1.79	56.04±1.92	45.91±1.83	42.79±2.08	53.13±1.86	52.21±1.73	50.81±1.84	44.29±2.25	30.85±2.36
	CPM [TPAMI 2022] (Zhang et al., 2022)	56.41±2.10	52.42±2.81	49.11±2.29	40.48±1.09	38.40±3.18	49.93±1.61	48.05±1.72	40.97±2.86	31.97±2.95	23.92±1.88
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	56.46±2.27	55.33±1.16	53.95±2.19	42.35±2.93	40.13±2.73	50.49±2.80	50.14±2.79	48.49±1.45	44.01±1.98	29.55±2.17
	PIMVC[TNNLS 2023] (Deng et al., 2023)	64.98±3.05	61.86±2.82	61.99±2.87	61.48±3.03	59.15±3.25	66.21±2.78	62.18±2.59	61.08±2.82	59.89±2.85	55.57±2.53
	GreatF [TCSVT 2023] (Wen et al., 2023a)	*65.02±3.81*	*62.91±2.49*	*62.86±2.72*	*61.93±2.99*	*59.87±3.51*	67.41± 2.93	*63.85±2.88*	*62.95±2.39*	*60.93±2.76*	*56.85±2.61*
	IHGAT	**66.08±2.33**	**65.95±2.11**	**65.41±1.75**	**64.91±2.30**	**62.73±2.81**	*67.56±2.11*	**66.39±2.35**	**66.05±2.21**	**62.01±3.10**	**60.89±2.85**
3Sources	SVD [1992] (Hotelling, 1992)	41.41±0.95	39.55±1.22	36.54±0.99	33.40±1.07	32.11±1.00	25.26±1.09	17.45±1.28	12.34±1.12	10.31±0.98	9.21±1.07
	Average [2010] (Enders, 2010)	44.53±2.13	39.47±2.44	36.35±2.16	34.45±2.11	29.43±3.26	26.24±2.92	18.38±3.35	12.29±1.80	9.32±2.75	7.61±1.28
	CRA [CVPR 2017] (Tran et al., 2017)	43.19±2.16	39.83±2.09	37.41±1.86	32.41±2.02	30.30±1.02	27.07±1.88	17.32±1.25	12.41±1.16	11.12±0.98	5.97±1.44
	iCmSC [TIP 2020] (Wang et al., 2020)	46.15±1.96	42.29±2.23	39.17±2.42	35.03±2.68	32.76±2.31	30.02±1.85	21.87±2.41	16.97±2.29	11.02±2.73	9.58±2.89
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	60.10±1.95	55.37±1.96	49.18±2.22	44.36±2.59	38.37±1.73	61.96±1.86	51.12±2.02	40.95±2.39	32.78±2.37	22.78±2.18
	GP-MVC [TIP 2021] (Wang et al., 2021)	47.90±1.60	44.81±1.92	41.66±1.75	36.09±2.01	33.88±2.35	31.05±1.72	22.96±2.07	18.27±2.16	12.77±2.37	10.81±2.61
	CPM [TPAMI 2022] (Zhang et al., 2022)	46.93±1.85	42.26±2.36	39.92±2.67	35.15±1.10	32.69±2.34	32.41±1.54	21.80±0.77	17.36±1.46	11.93±1.22	10.28±1.26
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	48.39±2.01	47.29±2.58	44.71±1.63	36.59±1.90	34.63±1.17	31.10±2.73	23.43±3.26	19.66±2.94	13.19±1.38	11.63±0.78
	PIMVC[TNNLS 2023] (Deng et al., 2023)	72.95±4.89	71.93±5.86	71.29±5.93	64.08±6.03	60.19±6.57	64.86±2.11	63.87±2.26	62.28±5.09	55.97±7.28	50.99±7.11
	GreatF [TCSVT 2023] (Wen et al., 2023a)	*74.20±5.45*	*73.29±6.72*	**73.14±6.56**	*68.36±6.89*	**66.27±6.86**	*65.08±1.09*	*64.08±1.88*	*62.96±7.13*	*60.81±8.01*	**59.29±7.03**
	IHGAT	**74.38±2.71**	**73.84±2.43**	*72.34*±**2.08**	**71.47±2.79**	*61.95±1.71*	**72.19±2.79**	**69.23±2.89**	**63.82±2.41**	**61.99±2.10**	*51.33±2.56*

The best results are in Bold, and the second best results are in Italics

Table 2 The clustering performance comparison on six datasets with high missing rates (ε)

Dataset	Method	ACC (%)		NMI (%)	
		ε=0.8	ε=0.9	ε=0.8	ε=0.9
CUB	iCmSC [TIP 2020] (Wang et al., 2020)	32.18±2.35	29.09±2.41	42.68±2.73	37.38±2.87
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	35.48±1.99	34.18±2.36	48.89±2.62	43.72±2.84
	GP-MVC [TIP 2021] (Wang et al., 2021)	32.96±2.09	30.92±2.38	44.29±2.34	40.72±2.59
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	35.67±1.75	34.59±1.95	48.98±1.82	46.85±1.98
	PIMVC[TNNLS 2023] (Deng et al., 2023)	*46.83±3.05*	*46.13±3.12*	*54.51±3.53*	*53.98±2.75*
	GreatF [TCSVT 2023] (Wen et al., 2023a)	45.15±3.28	44.87±3.57	52.95±3.11	52.53±2.92
	IHGAT	**53.92±2.96**	**53.23±2.89**	**60.09±2.58**	**58.86±2.64**
Football	iCmSC [TIP 2020] (Wang et al., 2020)	18.96±2.08	17.82±2.61	26.94±2.58	24.73±2.28
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	22.78±1.89	21.96±2.18	32.18±2.55	30.58±2.74
	GP-MVC [TIP 2021] (Wang et al., 2021)	20.87±1.96	19.71±2.42	28.48±2.73	26.66±2.47
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	24.18±1.78	23.79±1.96	39.57±1.59	34.82±1.91
	PIMVC[TNNLS 2023] (Deng et al., 2023)	*30.05±2.96*	*29.98±2.85*	*49.87±2.68*	*49.52±2.81*
	GreatF [TCSVT 2023] (Wen et al., 2023a)	28.18±2.83	27.97±2.91	48.31±2.58	47.76±2.74
	IHGAT	**33.89±2.39**	**32.16±2.43**	**54.22±2.39**	**52.24±1.98**
ORL	iCmSC [TIP 2020] (Wang et al., 2020)	24.28±2.35	22.89±2.27	41.74±2.38	39.94±2.19
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	32.82±2.83	29.96±2.44	51.86±2.72	49.82±2.74
	GP-MVC [TIP 2021] (Wang et al., 2021)	25.91±1.98	24.73±2.12	43.59±2.35	41.31±2.48
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	27.56±2.32	25.97±2.85	48.92±1.97	46.18±2.06
	PIMVC[TNNLS 2023] (Deng et al., 2023)	*47.89±3.05*	*47.01±3.12*	*75.01±3.22*	*74.87±2.63*
	GreatF [TCSVT 2023] (Wen et al., 2023a)	45.63±3.12	44.97±3.25	73.89±3.34	72.93±2.85
	IHGAT	**52.96±2.86**	**52.37±2.93**	**80.03±2.87**	**79.89±2.68**
PIE	iCmSC [TIP 2020] (Wang et al., 2020)	17.57±2.11	15.89±2.51	35.53±2.28	32.96±2.76
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	27.85±2.12	26.17±2.35	67.16±2.26	66.68±2.31
	GP-MVC [TIP 2021] (Wang et al., 2021)	18.07±2.28	17.15±2.36	36.08±2.45	35.41±2.38
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	20.12±2.46	18.86±2.73	39.75±2.61	36.87±2.57
	PIMVC[TNNLS 2023] (Deng et al., 2023)	*33.28±3.15*	*33.19±2.83*	*66.57±2.96*	*66.34±2.84*
	GreatF [TCSVT 2023] (Wen et al., 2023a)	32.89±2.96	32.46±2.69	65.99±3.07	65.12±2.98
	IHGAT	**37.53±2.53**	**37.21±2.72**	**72.96±2.82**	**72.59±2.65**
Politics	iCmSC [TIP 2020] (Wang et al., 2020)	27.17±2.13	23.93±1.92	12.02±1.88	10.81±2.15
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	31.79±2.13	28.96±2.32	14.63±2.27	12.57±2.61
	GP-MVC [TIP 2021] (Wang et al., 2021)	28.62±1.94	25.83±1.76	12.24±1.71	11.08±1.85
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	37.81±2.69	36.48±2.37	17.38±2.41	14.94±2.28
	PIMVC[TNNLS 2023] (Deng et al., 2023)	*47.82±2.76*	*45.39±2.49*	*23.57±2.82*	*15.98±2.76*
	GreatF [TCSVT 2023] (Wen et al., 2023a)	46.97±2.92	44.51±2.77	22.93±2.59	15.19±2.28
	IHGAT	**52.16±2.86**	**48.08±2.58**	**27.17±2.28**	**17.25±2.43**
3Sources	iCmSC [TIP 2020] (Wang et al., 2020)	23.18±2.57	18.92±2.68	6.37±1.83	5.05±2.03
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	30.89±2.77	28.67±2.78	9.86±2.01	8.27±2.15
	GP-MVC [TIP 2021] (Wang et al., 2021)	24.89±2.66	20.86±2.69	7.98±1.76	6.14±1.97
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	33.75±2.59	31.58±2.17	9.89±1.59	8.86±2.46
	PIMVC[TNNLS 2023] (Deng et al., 2023)	45.21±2.79	44.48±2.66	32.78±2.82	32.15±2.54
	GreatF [TCSVT 2023] (Wen et al., 2023a)	*48.92±2.83*	*48.11±2.93*	*35.97±2.73*	*34.76±2.62*
	IHGAT	**50.08±2.58**	**48.57±2.71**	**36.92±2.36**	**35.89±2.29**

The best results are in Bold, and the second best results are in Italics

Table 3 The clustering performance comparison on real-world missing data

Dataset	Method	ACC (%)	NMI (%)
ADNI	SVD [1992] (Hotelling, 1992)	38.27±0.98	0.95±0.24
	Average [2010] (Enders, 2010)	41.68±0.07	1.42±0.09
	CRA [CVPR 2017] (Tran et al., 2017)	37.71±1.16	0.42±0.11
	iCmSC [TIP 2020] (Wang et al., 2020)	39.96±2.02	2.35±1.18
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	41.85±1.89	4.17±1.27
	GP-MVC [TIP 2021] (Wang et al., 2021)	40.57±2.25	3.88±1.41
	CPM [TPAMI 2022] (Zhang et al., 2022)	42.39±0.87	4.68±0.39
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	43.98±0.75	5.42±2.51
	PIMVC[TNNLS 2023] (Deng et al., 2023)	*44.52±1.14*	*5.73±2.28*
	GreatF [TCSVT 2023] (Wen et al., 2023a)	44.29±1.38	5.66±2.71
	IHGAT	**46.96±2.15**	**5.98±2.96**
3Sources-partial	SVD [1992] (Hotelling, 1992)	21.79±0.96	2.26±0.38
	Average [2010] (Enders, 2010)	27.08±0.81	6.58±0.47
	CRA [CVPR 2017] (Tran et al., 2017)	40.05±1.38	27.75±2.64
	iCmSC [TIP 2020] (Wang et al., 2020)	48.16±2.54	45.96±2.37
	IMVTSC-MVI [AAAI 2021] (Wen et al., 2021)	54.99±2.04	54.37±1.72
	GP-MVC [TIP 2021] (Wang et al., 2021)	49.82±2.72	48.76±2.63
	CPM [TPAMI 2022] (Zhang et al., 2022)	52.08±3.12	41.97±2.54
	CPM-GAN [TPAMI 2022] (Zhang et al., 2022)	54.22±2.07	46.28±2.25
	PIMVC[TNNLS 2023] (Deng et al., 2023)	55.86±2.89	54.96±2.73
	GreatF [TCSVT 2023] (Wen et al., 2023a)	*56.31±2.51*	*55.72±2.48*
	IHGAT	**57.82±3.01**	**57.53±2.88**

The best results are in Bold, and the second best results are in Italics

experiments. Our algorithm was implemented in Torch 1.9.0 and carried all evaluations on a standard Ubuntu−16.04 system with NVIDIA 3090 Graphics Processing Units (GPUs). We set an initial learning rate of 0.02 on the other datasets except for the Football and PIE datasets, which had an initial learning rate of 0.01.

4.3 Baseline Methods

Eight baseline methods were used in the experiments, including SVD (Hotelling, 1992), Average (Enders, 2010), CRA (Tran et al., 2017), iCmSC (Wang et al., 2020), IMVTSC-MVI (Wen et al., 2021), GP-MVC (Wang et al., 2021), CPM (Zhang et al., 2022), CPM-GAN (Zhang et al., 2022), PIMVC (Deng et al., 2023), and GreatF (Wen et al., 2023a).

SVD (Hotelling, 1992): SVD is a matrix completion method by iterative soft thresholding of singular value decomposition.

Average (Enders, 2010): Average imputes missing parts with the average value of all samples in each modality.

CRA (Tran et al., 2017): CRA is composed of a set of stacked residual autoencoders, which can learn complex relationships among data from different modalities.

iCmSC (Wang et al., 2020): iCmSC is a novel incomplete cross-modal clustering method that integrates canonical correlation analysis and exclusive representation.

IMVTSC-MVI (Wen et al., 2021): IMVTSC-MVI incorporates the feature space based missing-modality inferring and manifold space based similarity graph learning into a unified framework.

GP-MVC (Wang et al., 2021): GP-MVC is a generative partial multi-modal clustering model with adaptive fusion and cycle consistency to solve the incomplete multi-modal problem by explicitly generating the data of missing modalities.

CPM (Zhang et al., 2022): CPM provides the comparative version of the CPM-Nets without the adversarial strategy.

CPM-GAN (Zhang et al., 2022): CPM-GAN can be regarded as generators. As for the discriminators, it uses the same structure as the generators. For the purpose of discrimination, a sigmoid layer is imposed on the output layer of each discriminator network.

PIMVC (Deng et al., 2023): PIMVC applies projection learning to IMmC, which solves the problem of information imbalance between different modalities.

GreatF (Wen et al., 2023a): GreatF provides an adaptive weighted matrix factorization model to obtain the representa-

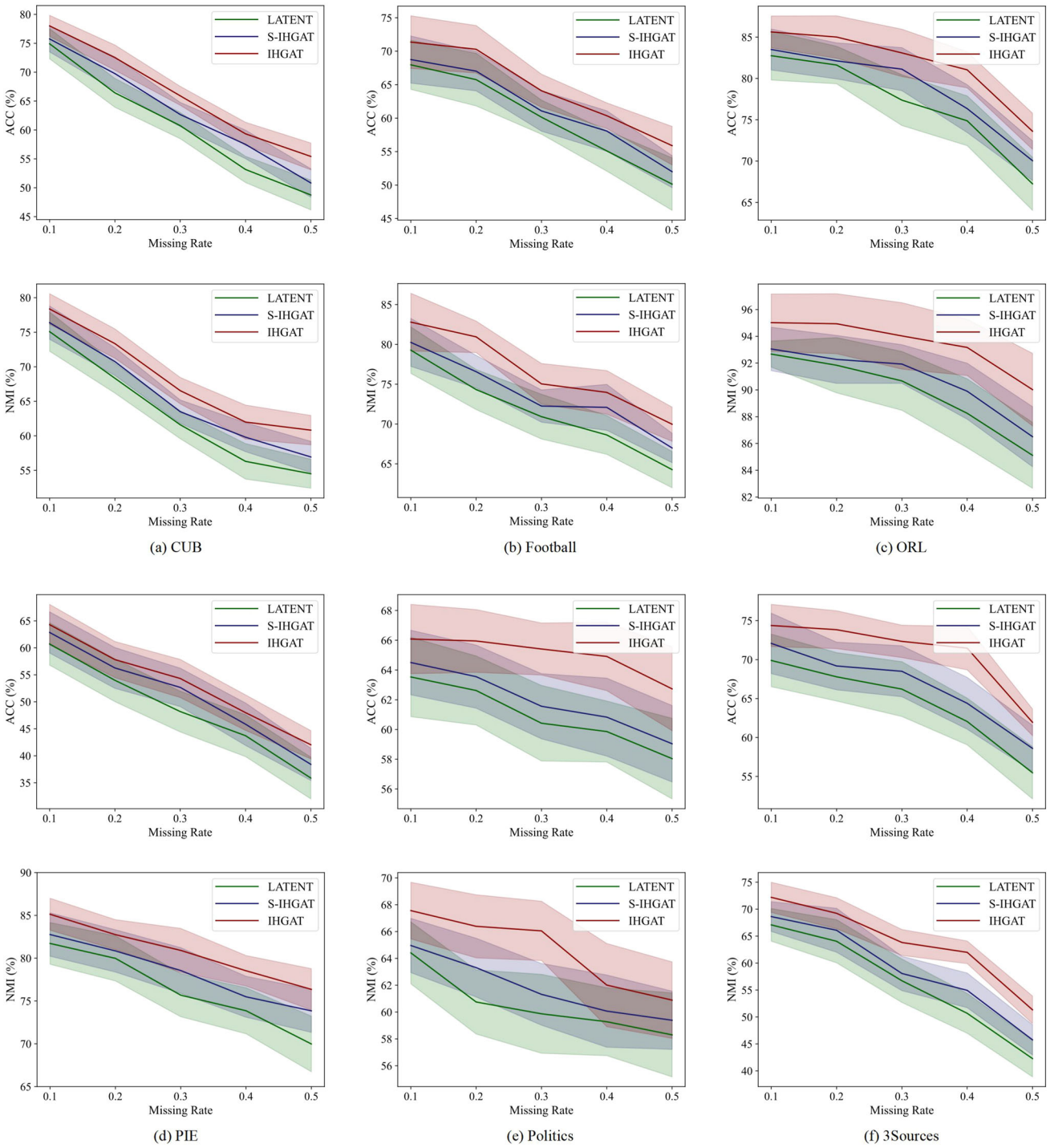

Fig. 2 Ablation study on (**a**) CUB, (**b**) Football, (**c**) ORL, (**d**) PIE, (**e**) Politics, and (**f**) 3Sources datasets

tion of every modality, which can enhance the weight of the discriminative features of all modalities for representation learning.

4.4 Incomplete Multi-Modal Clustering Performance

Experimental results are shown in Table 1. By analyzing the results, we have the following observations: (1) In terms of both ACC and NMI, our method achieves promising performance compared with all baselines. It performs the best

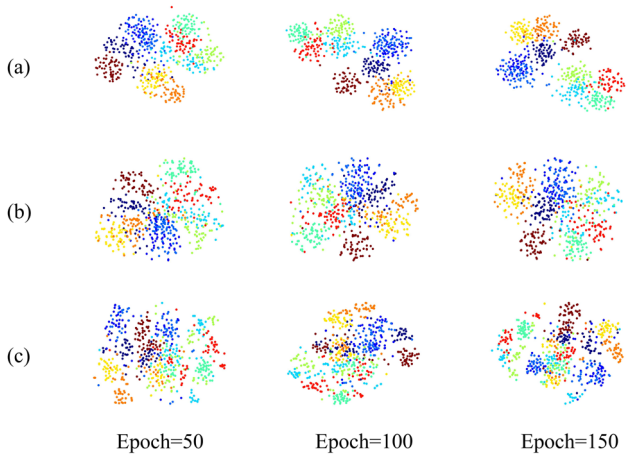

Fig. 3 Visualization of (**a**) IHGAT, (**b**) S-IHGAT, and (**c**) LATENT on CUB dataset with varying epochs

on most settings of all datasets in terms of both metrics, which validates the effectiveness of IHGAT. (2) Although the baselines can also achieve good performance at several low missing rates, a clear phenomenon of performance degeneration can be observed as the rate of missing data increases. Most existing baselines attempt to complete the missing modality, which may introduce extra noise when the missing situation is complex, or the missing rate is high. (3) On all missing rates, IHGAT generally achieves outstanding performance on all multi-modal datasets. On six datasets, we average the ACC and NMI of each method with different missing rates (ε from 0.1 to 0.5), and our method is 6.74% higher in ACC and 8.75% higher in NMI than the second-best method.

4.5 Multi-Modal Clustering Performance with High Missing Rates

Based on the above discussion, we have further analyzed the most state-of-the-art methods, including iCmSC, IMVTSC-MVI, GP-MVC, and CPM-GAN in Table 2. In terms of ACC, taking the missing rate (ε=0.9) for example, our method improves 18.25, 9.71, 25.4, 17.41, 14.35, and 16.33% over the second performers on CUB, Football, ORL, PIE, Politics, and 3Sources, respectively. IMVTSC-MVI hardly maintains good performance as the missing rate increases, but CPM-GAN is rather robust to modality-missing data. The missing modalities seriously affect the mining of information from multi-modal data. For a more comprehensive comparison, we compare the performance of IHGAT and the suboptimal method at high missing rates by taking the mean value. The suboptimal methods differ in different datasets, consisting of IMVTSC-MVI and CPM-GAN. The average ACC and NMI of IHGAT are 46.00 and 53.43%, respectively, and the suboptimal methods are 31.22 and 38.07%, respectively. IHGAT improves by 14.78 and 15.36% over the suboptimal method on ACC and NMI, respectively.

The combined analysis of Tables 1 and 2 shows that most baselines rely heavily on a large amount of paired multi-modal data by using shared information of latent representations to complete the missing modalities. When the missing modalities are large and complexly distributed, such methods may introduce additional noise and make it difficult to effectively complete the missing modalities. The above observations further validate the advantages of IHGAT. This suggests that it is beneficial to learn a more compact common representation for incomplete multi-modal clustering by considering both the structural information of missing data and the available information of non-missing data. It is clear that IHGAT is superior and more competitive, especially as the missing rate increases. This implies that our method can effectively explore the complex relationship between modalities and samples, even with a relatively large incomplete sample ratio.

4.6 Multi-Modal Clustering Performance on Real-World Missing Data

The existing IMmC is mostly based on publicly available multi-modal datasets, created by randomly removing portions of the data to form incomplete multi-modal datasets. In existing research, it is rare to encounter real-world incomplete multi-modal datasets. To further validate the effectiveness of IHGAT in real-world scenarios, we conducted experiments on two real-world incomplete multi-modal datasets, namely ADNI and 3Sources-partial.

As presented in Table 3, IHGAT still performs well on real-world missing datasets. In real-world scenarios, there are usually incomplete cases for multi-modal data. For instance, within medical applications, diverse subjects typically undergo various types of examinations. In the realm of web analysis, some websites encompass a variety of content, including text, images, and videos, while others may contain only a subset of these, resulting in data with missing modalities. As the number of modalities increases, the patterns of modality-missing, denoting the combinations of available modalities, become progressively intricate. Therefore, research on incomplete multi-modal data holds significant practical value and has a wide range of application scenarios.

4.7 Ablation Study

To verify the effectiveness of the similarity graph and modality-specific availability graphs, we visualized the representations on the CUB dataset to investigate the performance of IHGAT. Figure 3a, b, and c show the representations of IHGAT, S-IHGAT, and LATENT obtained in different

Table 4 The clustering performance comparison on three datasets with different modules

Dataset	Module	ACC (%)			NMI (%)		
		ε=0.1	ε=0.5	ε=0.9	ε=0.1	ε=0.5	ε=0.9
Football	w/o Similarity Graph and Availability Graph	67.81±2.77	50.69±3.96	25.95±3.62	78.96±3.86	64.19±2.53	36.02±3.93
	w/ Similarity Graph	68.92±3.73	52.47±2.52	29.11±3.73	80.17±3.28	66.83±2.37	39.18±3.51
	Integrated Heterogeneous Graph	**71.36±3.93**	**55.89±2.88**	**32.16±2.43**	**82.80±3.63**	**69.97±2.15**	**52.24±1.98**
	MLP	40.93±4.12	33.73±3.75	18.86±3.48	52.53±3.96	46.31±2.88	35.91±3.17
	GraphSAGE (Hamilton et al., 2017)	69.61±3.96	53.85±2.91	30.59±2.74	80.07±3.82	66.93±2.42	50.04±2.08
	GIN (Xu et al., 2019)	70.41±4.01	54.93±2.89	31.87±2.83	81.82±3.79	67.99±2.63	51.63±2.38
	GAT (Veličković et al., 2018)	**71.36±3.93**	**55.89±2.88**	**32.16±2.43**	**82.80±3.63**	**69.97±2.15**	**52.24±1.98**
	Frobenius Norm	70.07±3.57	54.24±2.97	31.07±2.88	80.79±3.71	68.11±2.85	50.81±2.52
	Kullback–Leibler Divergence	**71.36±3.93**	**55.89±2.88**	**32.16±2.43**	**82.80±3.63**	**69.97±2.15**	**52.24±1.98**
Politics	w/o Similarity Graph and Availability Graph	63.67±2.95	58.02±2.72	42.73±3.07	64.41±2.28	58.27±3.18	15.21±2.96
	w/ Similarity Graph	64.38±2.35	59.17±2.48	44.83±3.15	65.09±2.11	59.58±2.12	16.01±2.89
	Integrated Heterogeneous Graph	**66.08±2.33**	**62.73±2.81**	**48.08±2.58**	**67.56±2.11**	**60.89±2.85**	**17.25±2.43**
	MLP	56.72±3.96	40.71±3.71	24.59±3.45	50.29±3.18	30.31±2.75	10.06±3.41
	GraphSAGE (Hamilton et al., 2017)	64.21±2.54	60.08±2.99	45.63±2.72	64.28±2.48	58.13±2.97	15.09±2.69
	GIN (Xu et al., 2019)	65.39±2.49	61.24±2.87	46.38±2.63	65.37±2.51	59.38±2.98	16.89±2.74
	GAT (Veličković et al., 2018)	**66.08±2.33**	**62.73±2.81**	**48.08±2.58**	**67.56±2.11**	**60.89±2.85**	**17.25±2.43**
	Frobenius Norm	65.18±2.59	61.46±2.96	46.81±2.62	66.08±2.85	59.53±2.79	16.18±2.77
	Kullback–Leibler Divergence	**66.08±2.33**	**62.73±2.81**	**48.08±2.58**	**67.56±2.11**	**60.89±2.85**	**17.25±2.43**
3Sources	w/o Similarity Graph and Availability Graph	69.96±3.73	55.49±3.54	41.09±3.12	67.39±3.79	42.56±3.69	25.53±3.02
	w/ Similarity Graph	72.39±3.76	58.17±2.48	44.83±3.15	68.73±2.11	45.89±2.65	27.59±2.97
	Integrated Heterogeneous Graph	**74.38±2.71**	**61.95±1.71**	**48.57±2.71**	**72.19±2.79**	**51.33±2.56**	**35.89±2.29**
	MLP	45.86±3.85	31.94±3.43	19.12±3.67	28.77±3.38	8.51±3.27	6.11±3.28
	GraphSAGE (Hamilton et al., 2017)	73.96±2.75	60.17±2.81	47.81±2.69	70.96±2.85	50.01±2.62	34.19±2.49
	GIN (Xu et al., 2019)	73.07±2.93	60.19±2.78	46.99±2.83	70.18±2.76	50.12±2.81	34.02±2.57
	GAT (Veličković et al., 2018)	**74.38±2.71**	**61.95±1.71**	**48.57±2.71**	**72.19±2.79**	**51.33±2.56**	**35.89±2.29**
	Frobenius Norm	73.09±2.82	60.11±2.93	47.06±2.84	70.87±2.96	50.02±2.73	34.51±2.57
	Kullback–Leibler Divergence	**74.38±2.71**	**61.95±1.71**	**48.57±2.71**	**72.19±2.79**	**51.33±2.56**	**35.89±2.29**

The best results are in Bold

epochs. LATENT represents the proposed method without the similarity graph and modality-specific availability graphs; S-IHGAT uses the similarity graph only, and IHGAT uses both similarity graph and modality-specific availability graphs. As the number of epochs increases, the clusters of IHGAT are more compact, and the margins between different classes become more clear. It shows that the similarity graph and modality-specific availability graphs contribute substantially to the learning representation ability of IHGAT.

To further analyze the contribution of the similarity graph and modality-specific availability graphs in IHGAT, we conducted the ablation study with respect to the proposed method. As presented in Fig. 2, S-IHGAT substantially outperforms LATENT, which numerically indicates that it would be harmful to overlook the relationship between sample structures that can enhance multi-modal complementary information. Besides, IHGAT performs better than S-IHGAT, validating the effectiveness of the modality-specific availability graphs. Under the influence of the similarity graph and modality-specific availability graphs, IHGAT can indeed achieve better clustering results.

The graph adjacency matrix provides the possibility to correlate features and semantic representations, and the intrinsic structural information can be maintained to obtain more sufficient complementary information of different samples and modalities. The new features of a specific node are obtained by adding a nonlinear transformation to a weighted average of the neighboring features of the specific node in terms of their contribution. The new features are tighter and can further exploit complementarity. It can be intuitively seen through Figs. 2 and 3 that the similarity graph and modality-specific availability graphs have a significant impact on the performance of our proposed method, mainly because different constraints obtain different features.

As shown in Table 4, we conduct additional ablation experiments to explore the impact of different graph construction methods, attention mechanisms, and probability distribution. This will provide a deeper understanding of the contributions of these components. Additionally, we employed t-SNE visualization, as shown in Fig. 4, to illustrate that using multiple modalities to construct a unified representation results in more compact intra-class clusters and clearer inter-class boundaries. While directly aggregating encoder outputs into a common representation provides some improvement over single modality usage, utilizing multi-modal information to build a learnable unified latent representation yields superior overall performance. Our method reduces its heavy reliance on paired data by encoding multiple modalities into unified hidden representations. When the amount of data is large, we do not have to group the data like other works. The combination of the similarity graph and modality-specific availability graphs to form a set of integrated heterogeneous graphs can better explore the structural information of data and their relationship with each other.

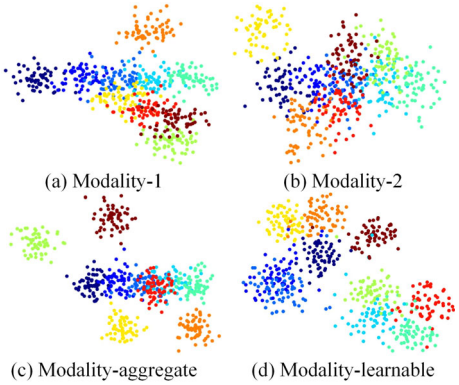

Fig. 4 Visualization of (**a**) Modality-1, (**b**) Modality-2, (**c**) Modality-aggregate, and (**d**) Modality-learnable on CUB

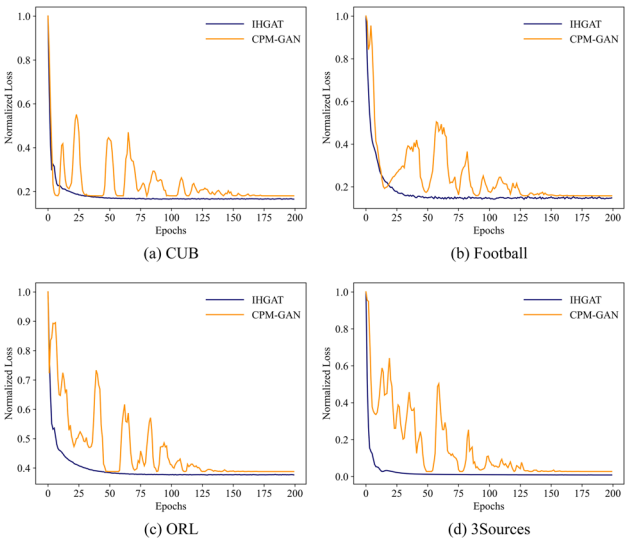

Fig. 5 Convergence analysis on (**a**) CUB, (**b**) Football, (**c**) ORL, and (**d**) 3Sources datasets

4.8 Convergence Analysis

To investigate the stability and convergence of the training process of IHGAT, we showed the convergence curves on multiple datasets ($\varepsilon = 0.9$) of IHGAT and CPM-GAN, a typical method of consistency strategy, respectively. As shown in Fig. 5, The training process of CPM-GAN is quite unstable on data with high missing rates. Consequently, the quality of the generator is challenging to control, which degrades the performance of the model significantly. Moreover, it also reveals the potential risk of introducing additional noise by the method of completing missing modalities with highly missing data. By contrast, IHGAT converges stably and fast in around 75 epochs, further demonstrating the performance advantage of IHGAT under complicated data distributions.

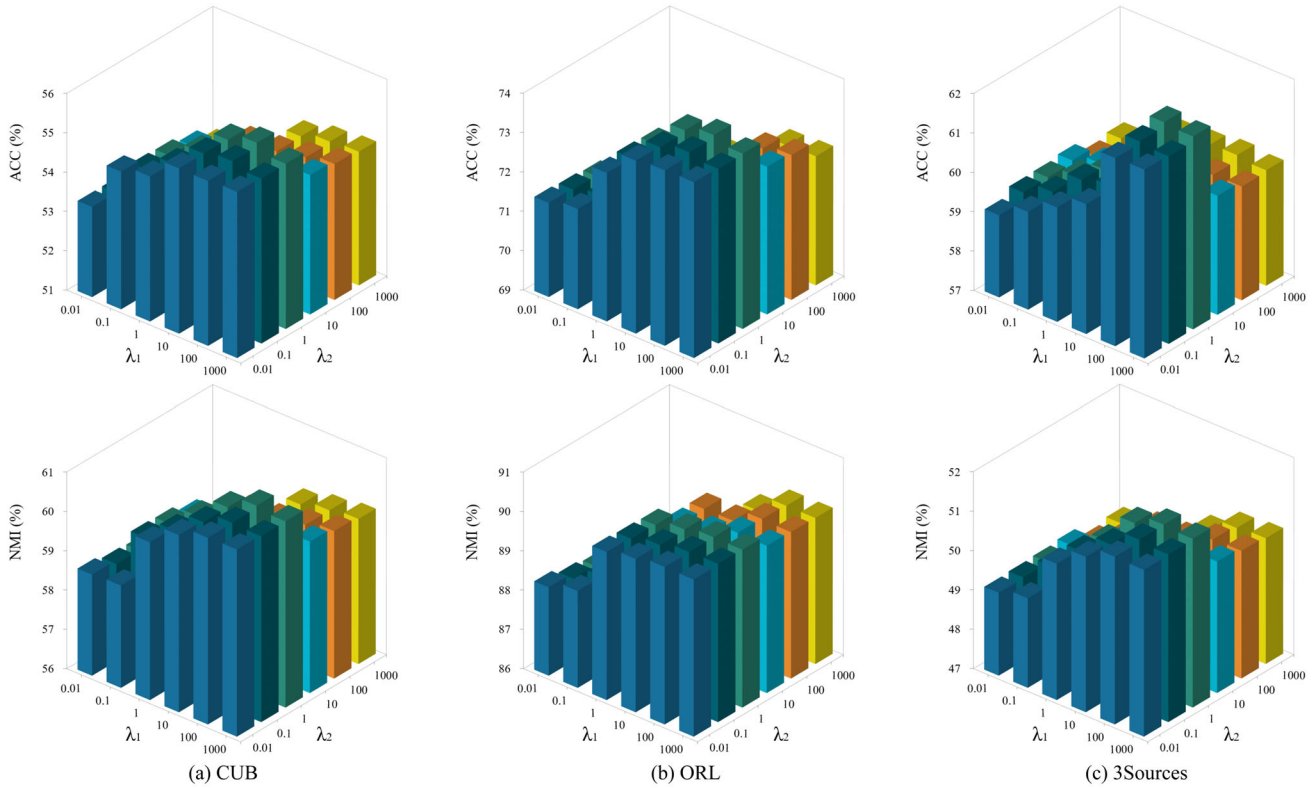

Fig. 6 Effect of the parameters λ_1 and λ_2 on (**a**) CUB, (**b**) ORL, and (**c**) 3Sources datasets

4.9 Parameter Analysis

Since λ_1 and λ_2 are the trade-off hyper-parameters that influence the clustering term and the reconstruction term in the final loss function, we analyzed the impacts of different values of λ_1 and λ_2 on the performance ($\varepsilon = 0.5$), and the results are shown in Fig. 6. It can be seen that IHGAT is not sensitive to λ_1 and λ_2, and the proposed method can reach a high level of performance whilst λ_1 is from the range of {10, 100, 1000} and λ_2 is from the range of {0.01, 0.1, 1}.

The number of nearest neighbors K and the dimensionality of the latent representations γ are the two main parameters of our method. In terms of K, since K nearest neighbors are used to obtain the similarity graph, we analyzed the influence of different K values on the proposed method. As shown in Fig. 7a, it can be observed that too small or large K values are adverse to the performance of the model. If the K is too small, it is easy to make the model complicated and thus overfitting. If the K value is too large, the result will be affected by distant points. Therefore, a medium K value, specifically $K = 5$, is appropriate for our method in the experiments.

In terms of γ, we visualized the influence of different values of γ on three datasets ($\varepsilon = 0.5$) in Fig. 7b, where the values of γ are ranged from {16, 32, 64, 128, 256}. Ten trials of experiments are conducted , and the average values of ACC and NMI are reported as the final results. According to Fig. 7b, it can be observed that different parameter settings greatly affect the performance of the method, and most datasets achieve better performance with γ in 64, which is by default a good choice.

5 Conclusion

In this paper, we proposed an effective method that deeply mines structural information to use complementarity information of different samples for IMmC. First, the similarity graph and modality-specific availability graphs are fused to form a set of integrated heterogeneous graphs. Thereafter, the attention mechanism is applied to the obtained integrated heterogeneous graphs to capture the complementarity information among different samples and modalities. In this way, complete representations can be learned for data with incomplete modalities. Finally, clustering is performed on the learned representations via embedding the consistency of probability distribution into the network. The proposed method does not require a large amount of paired data to model the missing modalities and shows significant improvements over the compared methods on six challenging benchmark datasets. More clear advantages of the proposed method over baselines can be observed with high missing rates of incomplete data.

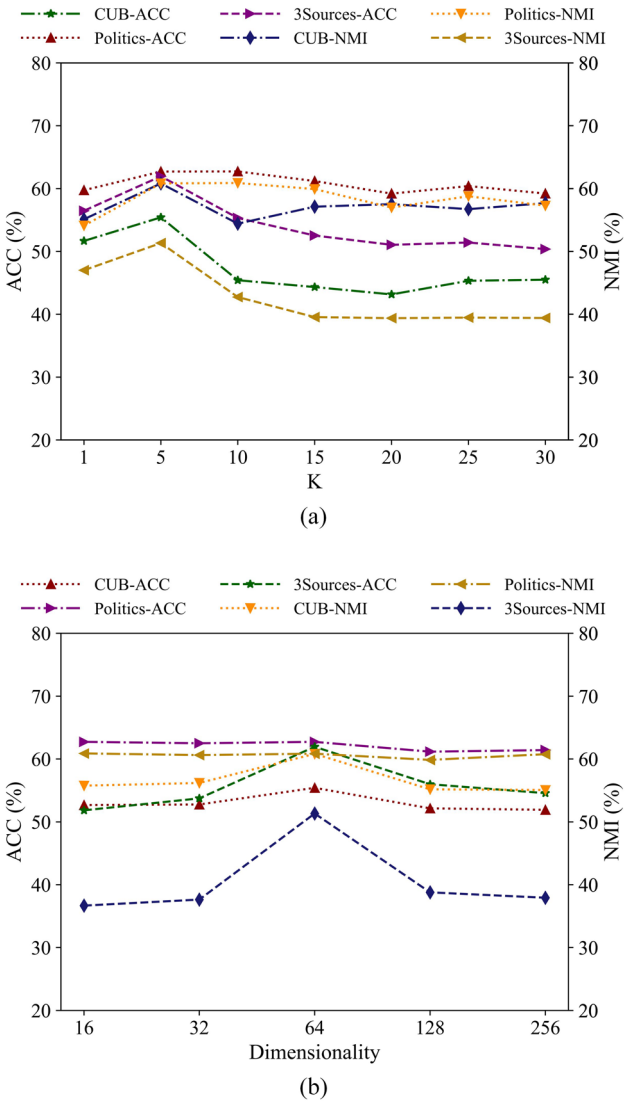

Fig. 7 Effect of the parameters (**a**) the number of nearest neighbor samples K and (**b**) the dimensionality of the latent representations in terms of ACC and NMI, respectively

Acknowledgements This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0116500, in part by the National Natural Science Foundation of China under Grants 62106174, 62222608, 62266035, and 61925602, and in part by Tianjin Natural Science Funds for Distinguished Young Scholar under Grant 23JCJQJC00270.

Funding National Science and Technology Major Project under Grant 2022ZD0116500; National Natural Science Foundation of China under Grants 62106174, 62222608, 62266035, and 61925602; Tianjin Natural Science Funds for Distinguished Young Scholar under Grant 23JCJQJC00270.

Data Availibility The CUB Wah et al. (2011) dataset can be obtained from https://www.vision.caltech.edu/datasets/cub_200_2011/. The Football dataset can be obtained from http://mlg.ucd.ie/aggregation/index.html. The ORL dataset can be obtained from https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. The PIE dataset can be obtained from http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html. The Politics dataset can be obtained from http://mlg.ucd.ie/aggregation/index.html. The 3Sources dataset can be obtained from http://mlg.ucd.ie/datasets/3sources.html.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443.

Bothorel, C., Cruz, J. D., Magnani, M., & Micenkova, B. (2015). Clustering attributed graphs: Models, measures and methods. *Network Science, 3*(3), 408–444.

Brasó, G., Cetintas, O., & Leal-Taixé, L. (2022). Multi-object tracking and segmentation via neural message passing. *International Journal of Computer Vision, 130*(12), 3035–3053.

Brissman, E., Johnander, J., Danelljan, M., & Felsberg, M. (2023). Recurrent graph neural networks for video instance segmentation. *International Journal of Computer Vision, 131*(2), 471–495.

Cao, Y., Luo, X., Yang, J., Cao, Y., & Yang, M. Y. (2022). Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection. *Information Fusion, 88*, 1–11.

Chang, S., Han, W., Tang, J., Qi, G.-J., Aggarwal, C. C., & Huang, T. S. (2015). Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 119–128.

Chen, L., Gao, Y., Huang, X., Jensen, C. S., & Zheng, B. (2020). Efficiently distributed clustering algorithms on star-schema heterogeneous graphs. *IEEE Transactions on Knowledge and Data Engineering,* pp. 1–15.

Chen, Y., Mancini, M., Zhu, X., & Akata, Z. (2022). Semi-supervised and unsupervised deep visual learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 1–23.

Chen, Y., Xiao, X., & Zhou, Y. (2019). Jointly learning kernel representation tensor and affinity matrix for multi-view clustering. *IEEE Transactions on Multimedia, 22*(8), 1985–1997.

Cheng, J., Wang, Q., Tao, Z., Xie, D.-Y., & Gao, Q. (2020). Multi-view attribute graph convolution networks for clustering. In *IJCAI*, pp. 2973–2979.

Deng, S., Wen, J., Liu, C., Yan, K., Xu, G., & Xu, Y. (2023). Projective incomplete multi-view clustering. *IEEE Transactions on Neural Networks and Learning Systems.*

Enders, C. K. (2010). *Applied missing data analysis.* Guilford press.

Fang, U., Li, M., Li, J., Gao, L., Jia, T., & Zhang, Y. (2023). A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering, 35*(12), 12350–12368.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing systems.*

Han, R., Gan, Y., Wang, L., Li, N., Feng, W., & Wang, S. (2023). Relating view directions of complementary-view mobile cameras via the human shadow. *International Journal of Computer Vision,* pp. 1–16.

Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in Statistics,* pp. 162–190. Springer.

Kumar, R., Chen, T., Hardt, M., Beymer, D., Brannon, K., & Syeda-Mahmood, T. (2013). Multiple kernel completion and its appli-

cation to cardiac disease discrimination. In *2013 IEEE 10th International Symposium on Biomedical Imaging,* pp. 764–767. IEEE.

Le, Q. & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning,* pp. 1188–1196. PMLR.

Li, L., Wan, Z., & He, H. (2021). Incomplete multi-view clustering with joint partition and graph learning. *IEEE Transactions on Knowledge and Data Engineering,* pp. 1–15.

Li, X., Wu, Y., Ester, M., Kao, B., Wang, X., & Zheng, Y. (2017). Semi-supervised clustering in attributed heterogeneous information networks. In *Proceedings of the 26th International Conference on World Wide Web,* pp. 1621–1629.

Lin, Y., Gou, Y., Liu, X., Bai, J., Lv, J., & Peng, X. (2023). Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(4), 4447–4461.

Lin, Y., Gou, Y., Liu, Z., Li, B., Lv, J., & Peng, X. (2021). Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 11174–11183.

Michieli, U., & Zanuttigh, P. (2022). Edge-aware graph matching network for part-based semantic segmentation. *International Journal of Computer Vision, 130*(11), 2797–2821.

Qi, G.-J., Aggarwal, C. C., & Huang, T. S. (2012). On clustering heterogeneous social media objects with outlier links. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining,* pp. 553–562.

Shao, W., He, L., & Philip, S. Y. (2015). Multiple incomplete views clustering via weighted nonnegative matrix factorization with $l_{2,1}$ regularization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases,* pp. 318–334. Springer.

Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y. (2016). A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering, 29*(1), 17–37.

Tao, Z., Liu, H., Li, J., Wang, Z., & Fu, Y. (2019). Adversarial graph embedding for ensemble clustering. In *International Joint Conferences on Artificial Intelligence Organization,* pp. 3562–3568.

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

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. *International Conference on Learning Representations*.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The caltech-ucsd birds-200-2011 dataset*.

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 (ICDM), pp. 1290–1295. IEEE.

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

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.

Wang, Q., Zhan, L., Thompson, P., & Zhou, J. (2020b). 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, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., & Yu, P. S. (2019). Heterogeneous graph attention network. In *The World Wide Web Conference,* pp. 2022–2032.

Wen, J., Xu, G., Tang, Z., Wang, W., Fei, L., & Xu, Y. (2023a). Graph regularized and feature aware matrix factorization for robust incomplete multi-view clustering. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wen, J., Yan, K., Zhang, Z., Xu, Y., Wang, J., Fei, L., & Zhang, B. (2020). Adaptive graph completion based incomplete multi-view clustering. *IEEE Transactions on Multimedia, 23*, 2493–2504.

Wen, J., Zhang, Z., Fei, L., Zhang, B., Xu, Y., Zhang, Z., & Li, J. (2023). A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 53*(2), 1136–1149.

Wen, J., Zhang, Z., Zhang, Z., Zhu, L., Fei, L., Zhang, B., & Xu, Y. (2021). Unified tensor framework for incomplete multi-view clustering and missing-view inferring. *In Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 10273–10281.

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.

Xie, D., Zhang, X., Gao, Q., Han, J., Xiao, S., & Gao, X. (2019). Multiview clustering by joint latent representation and similarity learning. *IEEE Transactions on Cybernetics, 50*(11), 4848–4854.

Xu, C., Tao, D., & Xu, C. (2015). Multi-view learning with incomplete views. *IEEE Transactions on Image Processing, 24*(12), 5812–5825.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? *International Conference on Learning Representations*.

Yang, L., Shen, C., Hu, Q., Jing, L., & Li, Y. (2019). Adaptive sample-level graph combination for partial multiview clustering. *IEEE Transactions on Image Processing, 29*, 2780–2794.

Yang, S., Li, L., Wang, S., Zhang, W., Huang, Q., & Tian, Q. (2019). Skeletonnet: A hybrid network with a skeleton-embedding process for multi-view image representation learning. *IEEE Transactions on Multimedia, 21*(11), 2916–2929.

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.

Zhan, K., Nie, F., Wang, J., & Yang, Y. (2018). Multiview consensus graph clustering. *IEEE Transactions on Image Processing, 28*(3), 1261–1270.

Zhang, C., Cui, Y., Han, Z., Zhou, J. T., Fu, H., & Hu, Q. (2022). Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*, 2402–2415.

Zhang, C., Fu, H., Hu, Q., Cao, X., Xie, Y., Tao, D., & Xu, D. (2018). Generalized latent multi-view subspace clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(1), 86–99.

Zhang, C., Fu, H., Wang, J., Li, W., Cao, X., & Hu, Q. (2020). Tensorized multi-view subspace representation learning. *International Journal of Computer Vision, 128*(8–9), 2344–2361.

Zhang, C., Song, D., Huang, C., Swami, A., & Chawla, N. V. (2019). Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* pp. 793–803.

Zhang, L., Zhao, Y., Zhu, Z., Shen, D., & Ji, S. (2018). Multi-view missing data completion. *IEEE Transactions on Knowledge and Data Engineering, 30*(7), 1296–1309.

Zhang, Y., Xiong, Y., Kong, X., Li, S., Mi, J., & Zhu, Y. (2018c). Deep collective classification in heterogeneous information networks. In *Proceedings of the 2018 World Wide Web Conference,* pp. 399–408.

Zhao, J., Wang, X., Shi, C., Hu, B., Song, G., & Ye, Y. (2021). Heterogeneous graph structure learning for graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence,* pp. 4697–4705.