Graph Embedded Contrastive Learning for Multi-View Clustering

Hongqing He[1,3] , **Jie Xu**[2,*] , **Guoqiu Wen**[1,3,*] , **Yazhou Ren**[4] , **Na Zhao**[2] , **Xiaofeng Zhu**[4,5]
[1]Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education,
Guangxi Normal University, Guilin 541004, China
[2]Singapore University of Technology and Design, Singapore 487372, Singapore
[3]Guangxi Key Lab of Multi-source Information Mining & Security, Guilin 541004, China
[4]School of Computer Science and Engineering,
University of Electronic Science and Technology of China, Chengdu 611731, China
[5]Hainan University, Haikou 570228, China

Abstract

Recently, numerous multi-view clustering (MVC) and multi-view graph clustering (MVGC) methods have been proposed. Despite significant progress, they still face two issues: I) MVC and MVGC are often developed independently for multi-view and multi-graph data. They have redundancy but lack a unified methodology to combine their strengths. II) Contrastive learning is usually adopted to explore the associations across multiple views. However, traditional contrastive losses ignore the neighbor relationship in multi-view scenarios and easily lead to false associations in sample pairs. To address these issues, we propose Graph Embedded Contrastive Learning for Multi-View Clustering. Concretely, we propose a process of view-specific pretraining with adaptive graph convolution to make our method compatible with both multi-view and multi-graph data, which aggregates the graph information into data and leverages autoencoders to learn view-specific representations. Furthermore, to explore the view-cross associations, we introduce the process of view-cross contrastive learning and clustering, where we propose the graph-guided contrastive learning that can generate global graph to mitigate the false association issue as well as the cluster-guided contrastive clustering for improving the model robustness. Finally, extensive experiments demonstrate that our method achieves superior performance on both MVC and MVGC tasks.

1 Introduction

The rich real-world applications have generated diverse formats of multi-view data, such as a sample often having multiple views, modalities, or graph structure information [Liang *et al.*, 2021; Chen *et al.*, 2024; Fu *et al.*, 2025]. Multi-View Clustering (MVC) can extract representations and discover patterns for multi-view data without relying on label information, which has attracted widespread attention from researchers [Bickel and Scheffer, 2004; Chaudhuri *et al.*, 2009; Nie *et al.*, 2017]. Recently, deep learning-based MVC has achieved significant progress and has been successfully applied in many fields like industry, internet, and medicine. Such deep MVC utilizes deep neural network models to improve the clustering ability and is conducive to exploring consistent and complementary information in multi-view data [Abavisani and Patel, 2018; Wen *et al.*, 2021; Luo *et al.*, 2024; Xu *et al.*, 2023; Xu *et al.*, 2024].

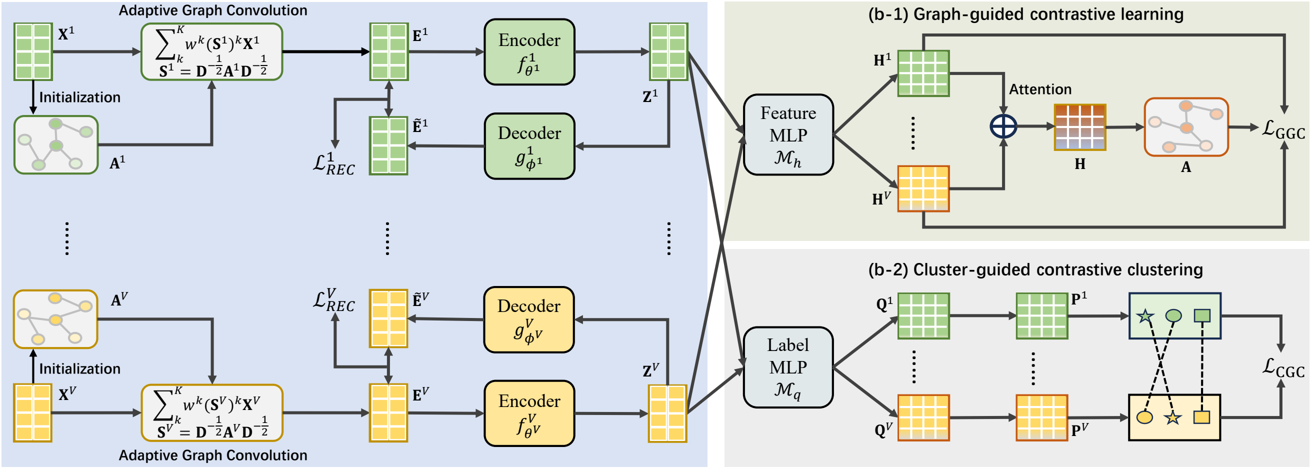
According to the data differences, we can finely divide existing deep MVC methods into two types: I) multi-view clustering for general multi-view data and II) multi-view graph clustering (MVGC) for multi-graph data, as follows.

For general multi-view data, each sample i contains multiple associated views $\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^V$, and these multi-view data often have different modalities, e.g., RGB and depth data are associated in DHA dataset [Lin *et al.*, 2012] and they are used for human action recognition and retrieval. Deep MVC typically combines deep models with view-cross clustering objectives to obtain discriminative representations. For example, [Abavisani and Patel, 2018] employed deep autoencoders to process multi-view/-modal data, and then conducted subspace clustering on the learned middle representations. [Perkins and Yang, 2019] combined deep encoder models with the alternating-view K-Means clustering and established a deep MVC method for dialogue intent induction. Recently, contrastive learning has gradually become a mainstream approach for deep MVC, which constructs multi-view data into positive/negative sample pairs to improve the representation discrimination for clustering. For instance, [Trosten *et al.*, 2021] proposed a contrastive deep MVC method that employs multiple encoders and minimizes contrastive loss to align the representations from different views. Furthermore, [Xu *et al.*, 2022] introduced contrastive learning into a multi-level feature learning framework that obtains final clustering predictions by averaging multiple network outputs.

For multi-graph data, the sample i owns its single-view data $\mathbf{x}_i$ and provides sample relation data in multiple graph structures $\mathbf{A}^1, \mathbf{A}^2, \ldots, \mathbf{A}^V$, e.g., ACM dataset [Jin *et al.*, 2021] describes the paper-author and paper-subject relation-

[*]Corresponding authors: Jie Xu (jiexuwork@outlook.com) and Guoqiu Wen (guoqiuwen@gxnu.edu.cn).

Code is available at https://github.com/SubmissionsIn/GMVC.


Adaptive Graph Convolution
$\mathbf{X}^1$
$\sum_k^K w^k(\mathbf{S}^1)^k\mathbf{X}^1$
$\mathbf{S}^1 = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}^1\mathbf{D}^{-\frac{1}{2}}$
Initialization
$\mathbf{A}^1$
$\mathbf{E}^1$
$\tilde{\mathbf{E}}^1$
$\mathcal{L}_{REC}^1$
Encoder $f_{\theta^1}^1$
Decoder $g_{\phi^1}^1$
$\mathbf{Z}^1$
$\mathbf{A}^V$
Initialization
$\mathbf{X}^V$
$\sum_k^K w^k(\mathbf{S}^V)^k\mathbf{X}^V$
$\mathbf{S}^V = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}^V\mathbf{D}^{-\frac{1}{2}}$
Adaptive Graph Convolution
$\mathcal{L}_{REC}^V$
$\tilde{\mathbf{E}}^V$
$\mathbf{E}^V$
Decoder $g_{\phi^V}^V$
Encoder $f_{\theta^V}^V$
$\mathbf{Z}^V$
(b-1) Graph-guided contrastive learning
Feature MLP $\mathcal{M}_h$
$\mathbf{H}^1$
$\mathbf{H}^V$
Attention
$\mathbf{H}$
$\mathbf{A}$
$\mathcal{L}_{\text{GGC}}$
(b-2) Cluster-guided contrastive clustering
Label MLP $\mathcal{M}_q$
$\mathbf{Q}^1$
$\mathbf{P}^1$
$\mathbf{Q}^V$
$\mathbf{P}^V$
$\mathcal{L}_{\text{CGC}}$

(a) View-specific pre-training with adaptive graph convolution

(b) View-cross contrastive learning and clustering

Figure 1: Our GMVC framework. (a) The view-specific pre-training with adaptive graph convolution organizes data input as multiple views and graphs. Then, it utilizes the adaptive graph convolution to aggregate graph structure information into the graph-embedded data $\mathbf{E}^v$, which is used to train the autoencoder with reconstruction loss $\mathcal{L}_{\mathbf{REC}}^v$ for refining the view-specific representation $\mathbf{Z}^v$. (b) The view-cross contrastive learning and clustering promote each other as follows. (b-1) Graph-guided contrastive learning achieves the interaction across multiple views to further refine $\{\mathbf{Z}^v\}_{v=1}^V$, which employs attention and graph optimization to get the global feature $\mathbf{H}$ and graph $\mathbf{A}$, and then $\mathbf{A}$ is integrated into our graph-guided contrastive loss $\mathcal{L}_{\mathbf{GGC}}$ to avoid false associations. (b-2) Cluster-guided contrastive clustering leverages comprehensive multi-view information in $\{\mathbf{Z}^v\}_{v=1}^V$ for end-to-end clustering, which learns clustering labels $\{\mathbf{Q}^v\}_{v=1}^V$ and cluster-strengthened labels $\{\mathbf{P}^v\}_{v=1}^V$, and then the cluster-guided contrastive loss $\mathcal{L}_{\mathbf{CGC}}$ is optimized for the clustering alignment across all views.

ships with two graphs. MVGC typically incorporates graph learning models into MVC frameworks to learn graph node representations, aiming to extract useful information from multi-graph data and then to facilitate clustering. For example, [Fan *et al.*, 2020] proposed a deep MVGC method that employs graph autoencoder models to learn the node representations shared by multi-graph data. [Wang *et al.*, 2022] proposed a multi-graph convolutional clustering network that fuses the representations from multiple graph encoders for clustering. Like contrastive learning in deep MVC, [Liu *et al.*, 2022] recently introduced contrastive learning into deep MVGC to improve the clustering effectiveness.

Although recent deep MVC and MVGC methods have achieved important progress, they face the following issues.

Firstly, current MVC and MVGC methods are individually designed, whereas their gap has not yet been bridged by a unified methodology and their commonalities result in redundancy. To be specific, general MVC methods cannot be applied to the clustering tasks of multi-graph data due to that their models are unable to accept the graph input. MVGC is technically similar to MVC and can be considered a subdomain of MVC [Chen *et al.*, 2022; Chen *et al.*, 2025]. Although MVGC methods are good at processing graph structure information in multi-graph data, experiments find that these graph data-specific methods have limited performance when dealing with general multi-view data (see Section 3.2). *Secondly*, contrastive learning typically treats data pairs in associated views as positive sample pairs, while data pairs without explicit association are considered as negative sample pairs [Trosten *et al.*, 2021]. This practice does not fully consider and utilize the neighbor relationships in multi-view scenarios and may have side effects. Concretely, multiple views of different samples might provide the information of the same class, but previous contrastive learning treats these data as negative sample pairs and increases the distance between their representations, causing the model to be negatively affected by their false associations. Although some methods [Chuang *et al.*, 2022; Sun *et al.*, 2024] tried to modify the contrastive loss to make it more robust, current multi-view methods have not yet considered incorporating graph structure information into the design of contrastive loss to alleviate this problem.

To address the aforementioned issues, this paper proposes GMVC: Graph Embedded Contrastive Learning for Multi-View Clustering as shown in Figure 1. First, GMVC has a view-specific pre-training process that can compatibly handle both multi-view and multi-graph data, where we propose to leverage the adaptive graph convolution and autoencoder to enhance the correlation among sample representations within each view. Then, GMVC performs the view-cross contrastive learning and clustering, where the graph-guided contrastive learning incorporates graph structure information into loss function for improving the discrimination of representations, and the cluster-guided contrastive clustering further explores the associations among multiple views for the robustness of clustering. In this way, GMVC can play a unified solution for clustering on both multi-view and multi-graph datasets.

Unlike existing deep MVC and MVGC methods, the main differences and contributions of this paper are as follows:

- We propose a novel deep MVC method, namely GMVC, which is equipped with a view-specific pre-training process and can leverage graph structure to refine representations. This makes our method applicable to both MVC and MVGC as well as combines their advantages.

- We design a novel contrastive loss that can utilize the graph structure of representations to guide the model in identifying more reliable positive and negative sample pairs, thereby promoting the effective usage of neighbor relationships in multi-view contrastive learning.
- We conduct extensive comparison experiments with state-of-the-art MVC and MVGC methods on multi-view and multi-graph datasets respectively, which verified the superiority of our proposed GMVC framework.

2 Method

Notations. In this paper, we define a multi-view dataset as $\{\mathbf{X}^v \in \mathbb{R}^{N\times D_v}\}_{v=1}^V$ including N samples with V view data, and define a multi-graph dataset as $\{\mathbf{X} \in \mathbb{R}^{N\times D}, \mathbf{A}^v \in \mathbb{R}^{N\times N}\}_{v=1}^V$ including N samples with V graph data. Assuming one dataset has different C classes, our goals of MVC and MVGC are the same to assign N samples to C clusters.

2.1 Preliminaries and our motivation

In this section, we present the preliminaries of deep MVC and MVGC in the literature and give our motivation.

For multi-view data, deep MVC methods usually adopt feedforward neural networks (e.g., autoencoder) to learn the data representation for clustering. Specifically, the learning paradigm of deep MVC can be illustrated as follows:

$$\mathcal{F}_{deep\ MVC} : \{\mathbf{X}^v\}_{v=1}^V \rightarrow \{\mathbf{Z}^v\}_{v=1}^V \rightarrow \mathbf{Q}. \tag{1}$$

Based on the learned representations $\{\mathbf{Z}^v\}_{v=1}^V$, some deep MVC methods [Zhou and Shen, 2020; Trosten *et al.*, 2021] focus on establishing effective fusion strategies to obtain one robust representation $\mathbf{Z}$ of multiple views and then produce final clustering result $\mathbf{Q}$. Some methods [Xu *et al.*, 2022; Tang and Liu, 2022; Chen *et al.*, 2023] choose to discover effective multi-view information interaction approaches to obtain individual robust clustering results $\{\mathbf{Q}^v\}_{v=1}^V$ from $\{\mathbf{Z}^v\}_{v=1}^V$ and then merge them into one result $\mathbf{Q}$.

For multi-graph data, deep MVGC methods leverage graph neural networks to extract the information in graphs, aiming to iteratively obtain node representations and refine graph structures for optimizing clustering. To be specific, the learning paradigm of deep MVGC can be illustrated as follows:

$$\mathcal{F}_{deep\ MVGC} : \mathbf{X}, \{\mathbf{A}^v\}_{v=1}^V \rightarrow \mathbf{Z} \rightarrow \mathbf{Q}. \tag{2}$$

Existing deep MVGC typically focus on designing different methods to effectively integrate multi-graph information in $\{\mathbf{A}^v\}_{v=1}^V$ into the representation learning for data $\mathbf{X}$, and the learned representation $\mathbf{Z}$ is then used to produce clustering result $\mathbf{Q}$. For example, [Fan *et al.*, 2020; Cheng *et al.*, 2021] train multiple graph decoders stacked on the representations to make them merge comprehensive information hidden in multi-graph data. [Ling *et al.*, 2023] further leverage pseudo labels to refine multiple graphs and then add them up to obtain a consensus graph and representation for clustering.

As shown in Eq. (1) and Eq. (2), the paradigm differences between deep MVC and MVGC make the researchers design their methods separately, causing the two incompatible and limiting their applications. To this end, our motivation is to propose a unified clustering framework for multi-view and multi-graph data, which combines the strengths of MVC and MVGC and hopefully provides a unified methodology for them. The learning paradigm of our method is expressed as:

$$\mathcal{F}_{ours} : \{\mathbf{X}^v\}_{v=1}^V\ or\ \mathbf{X}, \{\mathbf{A}^v\}_{v=1}^V \rightarrow \{\mathbf{Z}^v\}_{v=1}^V \rightarrow \mathbf{Q}. \tag{3}$$

To achieve this paradigm, we propose Graph Embedded Contrastive Learning for Multi-View Clustering (GMVC) as shown in Figure 1, which consists of the view-specific pre-training with adaptive graph convolution and view-cross contrastive learning & clustering that will be introduced below.

2.2 View-specific pre-training with adaptive graph convolution

This section introduces our proposed view-specific pre-training with adaptive graph convolution in GMVC, i.e., the process of $\{\mathbf{X}^v\}_{v=1}^V\ or\ \mathbf{X}, \{\mathbf{A}^v\}_{v=1}^V \rightarrow \{\mathbf{Z}^v\}_{v=1}^V$.

Data format unification. To make a unified framework compatible with both multi-view data and multi-graph data, we propose to organize them into a unified format. Specifically, our method GMVC achieves the following function:

$$\{\mathbf{X}^v\}_{v=1}^V\ or\ \mathbf{X}, \{\mathbf{A}^v\}_{v=1}^V \rightarrow \{\mathbf{X}^v, \mathbf{A}^v\}_{v=1}^V. \tag{4}$$

Firstly, if the input is multi-view data $\{\mathbf{X}^v\}_{v=1}^V$, we construct additional self-expression matrices $\{\mathbf{A}^v \in \mathbb{R}^{N\times N}\}_{v=1}^V$ for each view. Self-expression assumes that each sample can be reconstructed through a linear combination of other samples, that is, for any i-th sample in the v-th view $\mathbf{x}_i^v$, we have

$$\mathbf{x}_i^v = a_{i1}^v \mathbf{x}_1^v + a_{i2}^v \mathbf{x}_2^v + \cdots + a_{iN}^v \mathbf{x}_N^v, \tag{5}$$

where the coefficient $a_{ij}^v \in \mathbf{A}^v$ reflects the similarity between $\mathbf{x}_i^v$ and $\mathbf{x}_j^v$. In this way, the self-expression matrix $\mathbf{A}^v$ captures the neighbor information within each view, which can be used to produce the adjacency matrix as the graph data. Furthermore, considering that two samples with larger distances are likely to belong to the different classes [Zhou *et al.*, 2003], we optimize the following objective for each view:

$$\min_{\mathbf{A}^v} \|\mathbf{X}^v - \mathbf{A}^v\mathbf{X}^v\|_F^2 + \alpha \sum_{i,j=1}^N d_{ij}^v a_{ij}^v + \beta \sum_{i,j=1}^N (a_{ij}^v)^2, \tag{6}$$

where $d_{ij}^v = \left\|\mathbf{x}_i^v - \mathbf{x}_j^v\right\|_2^2$ is treated as a penalty term that encourages the self-expression matrix to focus on the local neighbor relationship. The third term prevents the matrix from reaching a trivial solution. α and β are non-negative parameters to trade off the three terms. Eq. (6) actually has the closed-form solution to obtain the optimal self-expression matrix $\mathbf{A}^v$ [Mo *et al.*, 2024]. Then, we leverage the following operation to make $\mathbf{A}^v$ become an adjacency matrix:

$$a_{ij} = \begin{cases} 1, & a_{ij} > 0\ . \\ 0, & a_{ij} \leq 0\ . \end{cases} \tag{7}$$

Secondly, if the input data is multi-graph data $\mathbf{X}, \{\mathbf{A}^v\}_{v=1}^V$, we can intuitively make V copies of $\mathbf{X}$ to form multiple views, i.e., $\{\mathbf{X}^v = \mathbf{X}\}_{v=1}^V$. Then, multi-view data and multi-graph data have the same format $\{\mathbf{X}^v, \mathbf{A}^v\}_{v=1}^V$, which allows us to develop the unified framework of MVC and MVGC.

View-specific pre-training. Given the data $\{\mathbf{X}^v, \mathbf{A}^v\}_{v=1}^V$, we further conduct the view-specific pre-training process which aims to achieve information aggregation between $\mathbf{X}^v$ and $\mathbf{A}^v$ and produces view-specific representations $\{\mathbf{Z}^v\}_{v=1}^V$.

Firstly, it is observed that the neighbor information in graphs can be embedded in data [Chanpuriya and Musco, 2022]. Therefore, we define the graph-embedded data $\mathbf{E}^v \in \mathbb{R}^{N\times d}$ and adopt the adaptive graph convolution (AGC) operations to aggregate the information in $\mathbf{X}^v$ and $\mathbf{A}^v$. Concretely, AGC computes the following tensor:

$$\mathcal{T}^v = \left[(\mathbf{S}^v)^0\mathbf{X}^v; (\mathbf{S}^v)^1\mathbf{X}^v; \ldots; (\mathbf{S}^v)^K\mathbf{X}^v\right] \in \mathbb{R}^{N\times d\times(K+1)}. \tag{8}$$

where $\mathbf{S}^v = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}^v\mathbf{D}^{-\frac{1}{2}}$ is the normalized adjacency matrix and $\mathbf{D} = \text{diag}(\mathbf{A}^v\mathbf{1})$ is the degree matrix. K controls the convolution orders. Then, for the i-th sample in the v-th view, we fetch $\mathbf{T}_i^v = \mathcal{T}^v[i,:,:] \in \mathbb{R}^{d\times(K+1)}$, and let $\mathbf{w}_i^v \in \mathbb{R}^{K+1}$ represent the importance of $(\mathbf{S}^v)^k\,\mathbf{X}^v$ at the k-th order. AGC optimizes $\mathbf{w}_i^v$ by minimizing the error of the least square:

$$\min_{\mathbf{w}_i^v} \left\|\mathbf{T}_i^v\mathbf{w}_i^v - (\mathbf{x}_i^v)^T\right\|_2^2. \tag{9}$$

After that, we compute the graph-embedded data $\mathbf{e}_i^v$ by

$$\mathbf{e}_i^v = \mathbf{T}_i^v\mathbf{w}_i^v \in \mathbf{E}^v. \tag{10}$$

Secondly, to further explore discriminative information in graph-embedded data $\mathbf{E}^v$, we leverage autoencoder [Hinton and Salakhutdinov, 2006] to learn view-specific representation $\mathbf{Z}^v$. Letting $f_{\theta^v}^v$ and $g_{\phi^v}^v$ denote the encoder and decoder of the v-th view, respectively, we optimize the following objective to pre-train the autoencoders for all views:

$$\mathcal{L}_{\mathbf{REC}} = \sum_{v=1}^V \left\|\mathbf{E}^v - g_{\phi^v}^v\left(f_{\theta^v}^v\left(\mathbf{E}^v\right)\right)\right\|_F^2, \tag{11}$$

where $\mathbf{Z}^v = f_{\theta^v}^v(\mathbf{E}^v) \in \mathbb{R}^{N\times d}$. By minimizing the reconstruction loss $\mathcal{L}_{\mathbf{REC}}$, the model is forced to learn the most discriminative information and filter out noise from data, which is conducive to discovering the clustering patterns.

Through Eqs. (10) and (11), the model aggregates the graph information into the graph-embedded data, and then produces the clustering-friendly view-specific representations. It is worth noting that the interaction across multiple views is not yet implemented, and the next section presents our view-cross contrastive learning and clustering in GMVC.

2.3 View-cross contrastive learning and clustering

This section introduces our view-cross contrastive learning and clustering in GMVC, i.e., the process of $\{\mathbf{Z}^v\}_{v=1}^V \rightarrow \mathbf{Q}$.

Graph-guided contrastive learning (GGC). To achieve the multi-view information interaction with contrastive learning for refining $\{\mathbf{Z}^v\}_{v=1}^V$, we construct a global feature $\mathbf{H} \in \mathbb{R}^{N\times f}$ and a global graph $\mathbf{A} \in \mathbb{R}^{N\times N}$, and propose GGC which makes the global graph information guide the contrastive learning to avoid its false associations. As shown in Figure 1, $\{\mathbf{Z}^v\}_{v=1}^V$ produce low-dimensional features $\{\mathbf{H}^v \in \mathbb{R}^{N\times f}\}_{v=1}^V$ by a multilayer perceptron (MLP) $\mathcal{M}_f$ and then obtain $\mathbf{H} \in \mathbb{R}^{N\times f}$ by

$$\mathbf{H} = \sum_{v=1}^V b^v\mathbf{H}^v = \sum_{v=1}^V b^v\mathcal{M}_f(\mathbf{Z}^v),\ b^v \in \mathbf{b}^v, \tag{12}$$

where $\mathbf{b}^v \in \mathbb{R}^v$ is a weighting vector to balance the importance among multiple views, and it is obtained through the following attention mechanism [Vaswani *et al.*, 2017]:

$$\mathbf{b}^v = \text{Softmax}(\text{Att}([\hat{\mathbf{h}}^1; \hat{\mathbf{h}}^2; \ldots; \hat{\mathbf{h}}^V])), \tag{13}$$

where $\hat{\mathbf{h}}^v$ is the mean value of the features $\mathbf{H}^v$ in the dimensions to simplify computation, i.e., $\hat{\mathbf{h}}^v = \frac{1}{f}\mathbf{H}^v\mathbf{1}$. Furthermore, the global graph $\mathbf{A}$ is inferred by the graph optimization as did in Eq. (6):

$$\min_{\mathbf{A}} \|\mathbf{H} - \mathbf{A}\mathbf{H}\|_F^2 + \alpha\sum_{i,j=1}^N d_{ij}a_{ij} + \beta\sum_{i,j=1}^N (a_{ij})^2. \tag{14}$$

$d_{ij} = \|\mathbf{h}_i - \mathbf{h}_j\|_2^2$ captures neighbor relationship among the global feature $\mathbf{H}$. Eq. (14) adopts a closed form solution similar to Eq. (6). In GGC, we mark the top n elements as 1 and mark others as 0 in each row of $\mathbf{A}$, to identify the reliable positive and negative sample pairs as the following rules:

$$\begin{cases} s^+ \in \mathcal{P}, & s^+ = \mathcal{D}(\mathbf{h}_i^m, \mathbf{h}_j^n) \text{ where } i = j, m \neq n. \\ \hat{s}^+ \in \hat{\mathcal{P}}, & \hat{s}^+ = \mathcal{D}(\mathbf{h}_i^m, \mathbf{h}_j^n) \text{ where } i \neq j, a_{ij} = 1. \\ s^- \in \mathcal{N}, & s^- = \mathcal{D}(\mathbf{h}_i^m, \mathbf{h}_j^n) \text{ where } i \neq j, a_{ij} = 0. \end{cases} \tag{15}$$

In the above rules, $\mathcal{P}$ denotes the positive sample pairs as did in previous contrastive learning. $\hat{\mathcal{P}}$ and $\mathcal{N}$ respectively denote the additional positive sample pairs and the rest negative sample pairs selected by our method. $\mathcal{D}(\mathbf{h}_i^m, \mathbf{h}_j^n)$ is the metric computed by cosine distance with a temperature, i.e., $1/\tau \cdot \langle\mathbf{h}_i^m, \mathbf{h}_j^n\rangle/\|\mathbf{h}_i^m\|_2\|\mathbf{h}_j^n\|_2$, $\mathbf{h}_i^m \in \mathbf{H}^m, \mathbf{h}_j^n \in \mathbf{H}^n$. Moreover, we define the graph-guided contrastive loss as follows:

$$\begin{split}\mathcal{L}_{GGC}^{m,n} = -\mathbb{E}_{s^+\in\mathcal{P}\cup\hat{s}^+\in\hat{\mathcal{P}}}[s^+ + \gamma\hat{s}^+ - \log(e^{s^+} + \\ \gamma e^{\hat{s}^+} + \textstyle\sum_{s^-\in\mathcal{N}} e^{s^-})].\end{split} \tag{16}$$

For any m, n-th views, minimizing our GGC loss $\mathcal{L}_{GGC}^{m,n}$ will enforce the model to explore their mutual information as well as avoid the possible false associations, thereby refining the view-specific representations to promote clustering. For all views, the total GGC loss is formulated as

$$\mathcal{L}_{\mathbf{GGC}} = \sum_{m=1}^V\sum_{n=m+1}^V \mathcal{L}_{GGC}^{m,n}. \tag{17}$$

Cluster-guided contrastive clustering (CGC). In order to fully utilize the comprehensive multi-view information in $\{\mathbf{Z}^v\}_{v=1}^V$ for clustering, we employ a MLP $\mathcal{M}_q$ to learn cluster labels $\{\mathbf{Q}^v \in \mathbb{R}^{N\times C}\}_{v=1}^V$ from $\{\mathbf{Z}^v\}_{v=1}^V$, and then we establish cluster-strengthened labels $\{\mathbf{P}^v \in \mathbb{R}^{N\times C}\}_{v=1}^V$ to make their cluster structures guide the view-cross clustering. Motivated by the usage of self-supervised learning in clustering [Xie *et al.*, 2016], we define $q_{ic}^v \in \mathbf{Q}^v$ as the probability that the i-th sample in the v-th view belongs to the c-th class, and compute the cluster-strengthened label $p_{ij}^v \in \mathbf{P}^v$ by

$$p_{ij}^v = \frac{(q_{ij}^v)^2/\sum_{i=1}^N q_{ij}^v}{\sum_{c=1}^C((q_{ic}^v)^2/\sum_{i=1}^N q_{ic}^v)}. \tag{18}$$

Datasets	DHA			NGs			WebKB			Caltech		
Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
K-Means	65.6 ± 2.9	79.8 ± 0.1	59.7 ± 2.7	20.6 ± 0.2	1.9 ± 0.3	21.0 ± 0.0	61.7 ± 0.8	0.2 ± 0.1	1.4 ± 0.0	$\underline{85.1}$ ± 0.1	$\underline{75.6}$ ± 0.1	$\underline{71.6}$ ± 0.2
DSIMVC	63.5 ± 4.6	77.8 ± 4.3	55.6 ± 4.7	63.0 ± 6.2	50.2 ± 5.9	43.9 ± 6.3	70.2 ± 1.4	25.0 ± 1.3	16.2 ± 2.3	49.8 ± 3.8	52.1 ± 3.2	38.1 ± 4.6
MFLVC	71.6 ± 1.1	81.2 ± 0.4	62.5 ± 0.7	90.8 ± 0.0	80.2 ± 0.0	79.2 ± 0.0	67.2 ± 2.1	24.5 ± 1.4	4.5 ± 2.1	75.2 ± 3.8	66.7 ± 2.6	58.6 ± 3.3
CPSPAN	66.3 ± 3.3	77.5 ± 1.0	62.7 ± 1.5	35.2 ± 0.2	21.5 ± 1.5	9.2 ± 0.5	77.1 ± 2.1	16.6 ± 4.2	12.5 ± 2.1	82.6 ± 4.0	73.2 ± 3.0	69.2 ± 4.2
CVCL	66.2 ± 6.3	75.4 ± 3.3	53.6 ± 6.3	56.8 ± 7.7	31.7 ± 7.8	28.1 ± 10.7	$\underline{74.1}$ ± 3.0	24.6 ± 2.6	$\underline{19.8}$ ± 3.3	80.3 ± 5.1	71.8 ± 3.8	65.8 ± 5.5
SCM	$\underline{80.4}$ ± 0.1	**84.0** ± 0.1	$\underline{70.0}$ ± 1.1	$\underline{96.5}$ ± 0.1	$\underline{89.3}$ ± 0.1	$\underline{91.4}$ ± 0.2	72.5 ± 2.4	$\underline{26.8}$ ± 5.2	15.5 ± 4.3	83.1 ± 0.2	71.0 ± 0.6	67.4 ± 0.6
GMVC [ours]	**82.5** ± 0.6	$\underline{83.5}$ ± 0.4	**70.4** ± 0.8	**97.3** ± 0.2	**92.0** ± 0.5	**93.3** ± 0.6	**80.2** ± 1.5	**32.6** ± 3.7	**34.9** ± 1.4	**87.9** ± 0.8	**79.6** ± 1.0	**75.8** ± 1.5

Table 1: Clustering performance of MVC methods on multi-view datasets. Bold and underline denote the best and the second-best results.

Eq. (18) encourages that large elements in $\mathbf{Q}^v$ become relatively larger in $\mathbf{P}^v$, and thus the cluster information with high-probability is strengthened. This cluster-strengthened label of one view can act as the supervised information to promote other views. To this end, we propose the CGC objective on $\{\mathbf{P}^v \in \mathbb{R}^{N\times C}\}_{v=1}^V$, which is different from the contrastive objective employed on $\{\mathbf{Q}^v \in \mathbb{R}^{N\times C}\}_{v=1}^V$ in other methods. Specifically, we define the positive and negative label pairs:

$$\begin{cases} l^+ \in \mathcal{P}', & l^+ = \mathcal{D}(\mathbf{p}^m_{.i}, \mathbf{p}^n_{.j}) \text{ where } i = j, m \neq n. \\ l^- \in \mathcal{N}', & l^- = \mathcal{D}(\mathbf{p}^m_{.i}, \mathbf{p}^n_{.j}) \text{ where } i \neq j. \end{cases} \tag{19}$$

For any m, n-th views, the CGC loss is further defined as

$$\mathcal{L}^{m,n}_{CGC} = -\mathbb{E}_{l^+\in\mathcal{P}'}[l^+ - \log(e^{l^+} + \sum\nolimits_{l^-\in\mathcal{N}'} e^{l^-})], \tag{20}$$

and minimizing it will achieve the view-cross agreement between positive label pairs as well as disagreement between negative label pairs, which is consistent with the intuition that the same clusters in different views should have the consistent cluster labels. For all views, the CGC loss is formulated as

$$\mathcal{L}_{\mathbf{CGC}} = \sum_{m=1}^{V}\sum_{n=m+1}^{V} \mathcal{L}^{m,n}_{CGC} + \sum_{v=1}^{V}\sum_{c=1}^{C} r^v_c \log r^v_c, \tag{21}$$

where the second term is a commonly-used regularization for avoiding trivial solutions and it has $r^v_c = \frac{1}{N}\sum_{i=1}^{N} p^v_{ic}$.

Objective function. Our proposed framework conduct the graph-guided contrastive learning and the cluster-guided contrastive clustering simultaneously. Overall, the training loss consists of the following three components:

$$\mathcal{L} = \mathcal{L}_{\mathbf{REC}} + \mathcal{L}_{\mathbf{GGC}} + \mathcal{L}_{\mathbf{CGC}}. \tag{22}$$

where $\mathcal{L}_{\mathbf{REC}}$ constrains the view-specific representations $\{\mathbf{Z}^v\}_{v=1}^V$ to maintain the discriminative information of data. Minimizing $\mathcal{L}_{\mathbf{GGC}}$ achieves multi-view information interaction to refine $\{\mathbf{Z}^v\}_{v=1}^V$. Meanwhile, minimizing $\mathcal{L}_{\mathbf{GGC}}$ learns semantic consistent clustering labels $\{\mathbf{Q}^v\}_{v=1}^V$ from $\{\mathbf{Z}^v\}_{v=1}^V$. Finally, we calculate the average of the cluster labels of all views to obtain the final clustering results by

$$q_{ij} = \frac{1}{V}\sum_{v=1}^{V} q^v_{ij},\ q_{ij} \in \mathbf{Q}. \tag{23}$$

Eq. (23) finishes the process of $\{\mathbf{Z}^v\}_{v=1}^V \rightarrow \{\mathbf{Q}^v\}_{v=1}^V \rightarrow \mathbf{Q}$, and the hard cluster label is $y_i = \arg\max_c q_{ic}, c \in \{1, 2, \ldots, C\}$. In this way, our proposed method can obtain end-to-end clustering results for input data, which is compatible with both multi-view and multi-graph datasets.

3 Experiments

3.1 Experimental setup

In this subsection, we briefly present the datasets, comparison methods, and evaluation protocol. More information can be found in the supplementary materials of our GMVC code.

Datasets. We conduct experiments on 8 public benchmarks, including 4 multi-view datasets, i.e., DHA [Lin *et al.*, 2012], NGs [Hussain *et al.*, 2010], WebKB [Sun *et al.*, 2007], Caltech [Fei-Fei *et al.*, 2004], and 4 multi-graph datasets, i.e., ACM [Jin *et al.*, 2021], IMDB [Jin *et al.*, 2021], Texas [Ling *et al.*, 2023], Chameleon [Ling *et al.*, 2023].

Comparison methods. In this paper, the baseline methods include 2 classical single-view clustering methods (i.e., K-Means [MacQueen, 1967], VGAE [Kipf and Welling, 2016]) and 5 recent MVC methods (i.e., DSIMVC [Tang and Liu, 2022], MFLVC [Xu *et al.*, 2022], CVCL [Chen *et al.*, 2023], CPSPAN [Jin *et al.*, 2023], SCM [Luo *et al.*, 2024]) and 5 recent MVGC methods (i.e., O2MAC [Fan *et al.*, 2020], MvAGC [Lin and Kang, 2021], MVGC [Xia *et al.*, 2022], DuaLGR [Ling *et al.*, 2023], BTGF [Qian *et al.*, 2024]).

Evaluation. We leverage three metrics for comprehensive evaluation, i.e., clustering accuracy (ACC), normalized mutual information (NMI), adjusted rand index (ARI), and report the mean results with standard deviation of 10 runs.

3.2 Comparison experiments

Tables 1, 2, 4 demonstrate the effectiveness of our GMVC compared with MVC and MVGC methods, respectively.

MVC on multi-view datasets. When our GMVC conducts clustering tasks on multi-view datasets as shown in Table 1, we have the following observations: 1) Our GMVC is compatible with MVC tasks and achieves better performance than the comparison methods. For example, on WebKB, our GMVC obtains 18.5% improvement from the single-view method K-Means, and surpasses the recent deep MVC method SCM by 7.7% in ACC values. This indicates that our GMVC can effectively explore the multi-view information for clustering. 2) Our GMVC of integrating graph structure information into contrastive learning is conducive to improving the model robustness. For example, the deep MVC methods CPSPAN, and CVCL also use contrastive learning, and they perform well on WebKB, Caltech but underperform on DHA, NGs. On these multi-view datasets, our GMVC consistently achieves the robust clustering performance.

MVGC on multi-graph datasets. Table 2 shows the comparison of our GMVC conducting clustering tasks on multi-graph datasets. For further comparison, Table 4 reports the

Datasets	ACM			Chameleon			IMDB			Texas		
Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
VGAE	44.4 ± 0.6	8.3 ± 0.7	5.0 ± 0.4	35.4 ± 1.0	15.1 ± 0.7	12.4 ± 0.6	41.2 ± 1.0	1.8 ± 0.5	0.7 ± 0.4	55.3 ± 1.8	12.7 ± 4.4	21.7 ± 8.4
O2MAC	89.7 ± 0.5	67.2 ± 1.1	72.2 ± 1.3	33.5 ± 0.3	12.3 ± 0.7	8.9 ± 1.2	42.9 ± 2.5	3.1 ± 1.1	3.8 ± 1.5	46.7 ± 2.4	8.7 ± 0.8	14.6 ± 1.8
MvAGC	84.4 ± 0.9	57.8 ± 1.3	60.1 ± 1.8	29.2 ± 0.9	10.8 ± 0.8	3.3 ± 1.7	**56.3** ± 0.0	3.7 ± 0.0	**9.4** ± 0.0	54,3 ± 2.6	5.4 ± 2.8	1.1 ± 4.1
MVGC	80.3 ± 2.0	57.6 ± 2.3	52.5 ± 3.4	32.8 ± 0.4	12.6 ± 0.3	5.1 ± 0.6	41.7 ± 2.8	0.2 ± 0.0	0.1 ± 0.0	41.8 ± 2.6	8.1 ± 3.3	7.8 ± 3.1
DuaLGR	92.4 ± 0.2	72.5 ± 0.5	78.8 ± 0.5	42.2 ± 0.2	18.5 ± 0.1	13.6 ± 0.1	44.3 ± 3.2	4.5 ± 2.0	5.2 ± 2.5	55.4 ± 2.1	33.6 ± 5.1	24.1 ± 3.7
BTGF	90.2 ± 1.4	69.3 ± 2.6	73.6 ± 3.2	35.8 ± 0.0	17.2 ± 0.0	11.5 ± 0.0	52.7 ± 5.5	4.6 ± 2.4	7.3 ± 5.8	58.5 ± 0.0	22.7 ± 0.0	20.5 ± 0.0
GMVC [ours]	**93.4** ± 0.1	**76.2** ± 0.2	**81.3** ± 0.2	**42.3** ± 0.3	**19.3** ± 0.5	**14.4** ± 0.4	47.7 ± 0.3	**9.1** ± 0.1	**9.4** ± 0.2	**60.1** ± 0.1	**36.4** ± 0.3	**31.0** ± 0.3

Table 2: Clustering performance of MVGC methods on multi-graph datasets. Bold and underline denote the best and the second-best results.

Datasets	DHA			NGs			WebKB			Caltech			ACM			IMDB			Texas			Chameleon		
Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
(1) w/o AGC	81.1	83.1	68.7	93.7	83.5	85.0	77.6	31.9	29.9	86.3	76.8	72.8	86.0	55.1	62.7	40.5	3.4	3.8	58.4	**38.5**	**32.3**	39.8	18.0	13.0
(2) w/o GGC	77.0	81.3	65.5	82.2	70.1	65.7	73.9	28.6	22.7	84.5	74.7	70.3	92.9	74.5	80.1	46.9	7.6	8.3	54.1	37.6	24.9	38.9	14.8	10.8
(3) w/o CGC	75.4	80.9	63.6	94.8	85.9	87.6	78.2	0.5	0.5	44.9	39.5	24.6	53.3	13.4	10.5	42.3	5.5	3.8	48.6	34.2	23.7	35.3	13.3	8.9
(4) w/ normal CL	82.3	**83.5**	70.1	**97.4**	**92.4**	**93.8**	71.4	24.4	18.4	84.6	75.2	71.0	93.3	76.1	81.2	45.8	7.4	7.9	58.4	34.3	27.4	40.7	17.1	12.4
(5) w/ normal CC	69.6	78.5	57.9	95.6	87.5	89.4	80.1	**34.9**	**35.5**	79.8	72.2	64.8	92.9	75.3	80.2	46.6	7.9	8.4	58.4	33.3	28.2	35.3	14.1	10.0
(6) **GMVC**	**82.5**	**83.5**	**70.4**	97.3	92.0	93.3	**80.2**	32.6	34.9	**87.9**	**79.6**	**75.8**	**93.4**	**76.2**	**81.3**	**47.7**	**9.1**	**9.4**	**60.1**	36.4	31.0	**42.3**	**19.3**	**14.4**

Table 3: Clustering performance of the proposed GMVC framework using different components on multi-view and multi-graph datasets.

Datasets	DHA		
Metrics	ACC	NMI	ARI
VGAE	7.5 ± 0.1	3.4 ± 0.5	2.1 ± 0.0
O2MAC	13.5 ± 0.6	18.7 ± 1.0	1.2 ± 0.3
MvAGC	50.6 ± 1.4	61.6 ±1.4	31.5 ± 2.2
MVGC	12.5 ± 0.7	16.6 ± 0.8	-0.3 ± 0.2
DuaLGR	65.6 ± 1.2	76.0 ± 0.3	52.6 ± 0.8
BTGF	47.2 ± 3.1	64.1 ± 2.4	34.0 ± 3.8
GMVC [ours]	**82.5** ± 0.6	**83.5** ± 0.4	**70.4** ± 0.8

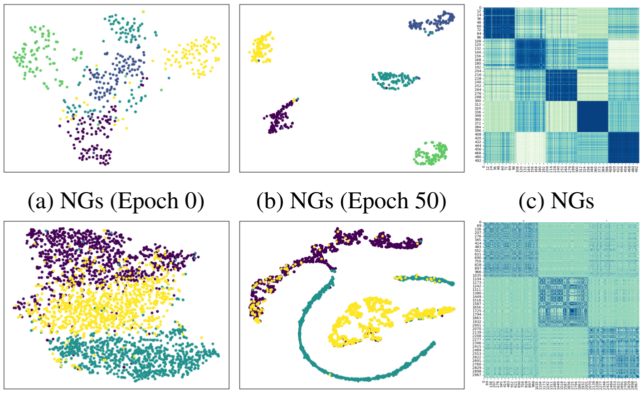
Table 4: Clustering performance of MVGC methods on DHA.

performance of MVGC methods on non-graph data. From the results, we have the following conclusions: 1) Our GMVC is compatible with MVGC tasks and achieves superior or comparable performance for the comparison methods. For example, on ACM, our GMVC obtains 49% improvement from the single-view method VGAE, and surpasses the recent deep MVGC method BTGF by 3.2% in ACC values. This indicates that our GMVC can effectively explore the multi-graph information for clustering. 2) Previous MVGC methods usually are incompatible with MVC tasks whereas our GMVC combines the advantages of MVC and MVGC. For instance, we construct multi-graph input on DHA by the same way in our GMVC, and then use MVGC methods to conduct clustering as shown in Table 4. The results indicate that MVGC methods might have very limited performance on non-graph data, and thus our proposed unified framework of MVC and MVGC could own wider application potential.

3.3 Ablation study

Table 3 displays the ablation studies for verifying the importance of each component in our GMVC framework.

Specifically, Item (1) “w/o AGC” denotes the variant of GMVC which does not adopt the view-specific pre-training with adaptive graph convolution, and it directly feeds the raw data into the subsequent modules for multi-view clustering. Item (2) “w/o GGC” is the framework without using the graph-guided contrastive learning, and clustering performance is achieved through training the CGC module. Item (3) “w/o CGC” is the framework without using the cluster-guided contrastive clustering. Item (4) “w/ normal CL” denotes the framework with the normal contrastive loss for multi-view representation learning. Item (5) “w/ normal CC” is the framework with the normal contrastive clustering loss. Compared with our complete components, i.e., Item (6) GMVC, the above five variants are respectively with incomplete components and obtain suboptimal results on both multi-view datasets and multi-graph datasets, thereby demonstrating the importance of each component in our GMVC framework.



(a) NGs (Epoch 0) (b) NGs (Epoch 50) (c) NGs

(d) ACM (Epoch 0) (e) ACM (Epoch 50) (f) ACM

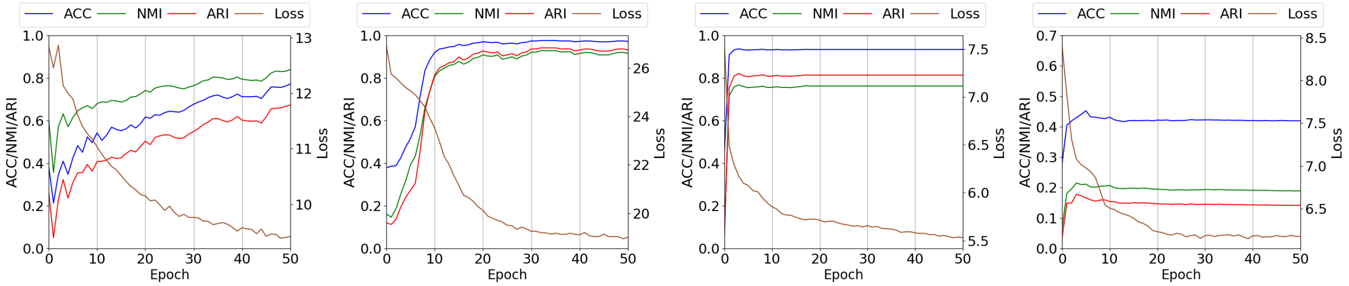
Figure 2: Visualization of global features and their correlation graphs for NGs (a,b,c) and ACM (d,e,f) during the learning process.

4 Model Analysis

In this part, we present the visualization and parameter analysis of our GMVC framework to understand its behaviors.

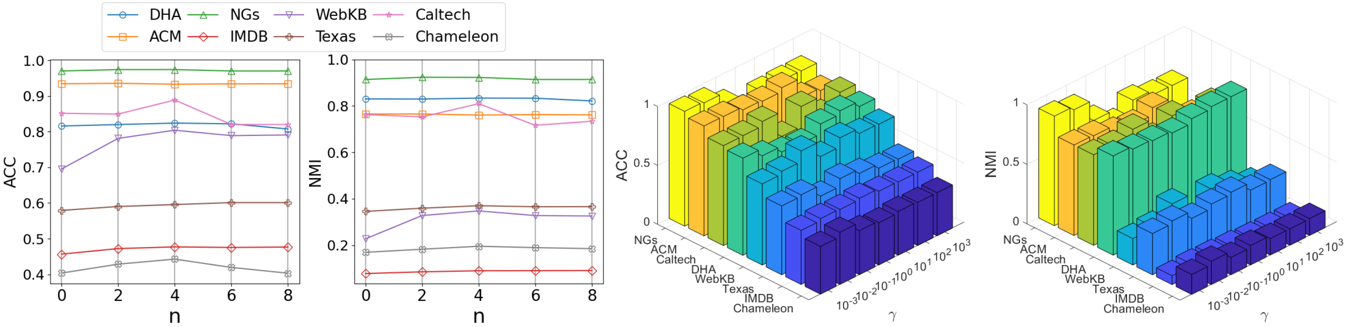
4.1 Multi-view and multi-graph representations

We take the multi-view dataset NGs and multi-graph dataset ACM as examples and visualize their representation learning processes by GMVC. In Figures 2(a,b,d,e), the global fea-



(a) DHA (b) NGs (c) ACM (d) Chameleon

Figure 3: The training loss curve and clustering performance of GMVC on multi-view datasets (a,b) and multi-graph datasets (c,d).



(a) ACC (Left) and NMI (Right) $vs.$ GGC parameter n (b) ACC $vs.$ GGC parameter γ (c) NMI $vs.$ GGC parameter γ

Figure 4: Clustering performance with different settings of GGC parameters on four multi-view datasets and four multi-graph datasets.

tures with the same class are plotted in the same color by t-SNE. By comparing Figures 2(a,b) and Figures 2(d,e), we can observe that the global features become more discriminative from Epoch 0 to Epoch 50. The correlation graphs on the global features also exhibit clear block diagonal structures as shown in Figures 2(c,f), which indicates that GMVC has captured the true class information from multiple views.

4.2 Training loss and performance

In Figure 3, we visualize the curves of loss as well as clustering performance during the model training process of GMVC. It can be found that the loss curves of GMVC have the smooth downward trend on both multi-view datasets and multi-graph datasets. Additionally, the steady increase of clustering performance curve (i.e., ACC, NMI, ARI) further validates that the model gradually learns the correct cluster structures of these datasets. This indicates that the model has good convergence and compatibility for the two kinds of data.

4.3 Parameter analysis

In our GMVC framework, we employ the non-negative parameters n and γ to control the GGC loss as shown in Eq. (16), where n selects the top n elements in each row of $\mathbf{A}$ to identify the additional positive sample pairs, and γ balances the contrastive losses between the original and additional positive sample pairs. In order to investigate the effects of n and γ, we conduct the parameter analysis in Figure 4. Specifically, Figure 4(a) displays n in the range of $[0, 2, 4, 6, 8]$, and the corresponding clustering performance is insensitive. In our experiments, we set $n = 4$ for all multi-view datasets and set $n = 6$ for all multi-graph datasets. Figures 4(b,c) display γ within the range of $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]$, where we find that γ has the stability of clustering performance on the most of all multi-view and multi-graph datasets. In our experiments, γ is set within the range of $[10^{-3}, 10^{-1}]$.

5 Conclusion

In existing research, Multi-View Clustering (MVC) and Multi-View Graph Clustering (MVGC) are often studied separately, which prevents the advantages of these two approaches from being fully utilized and limits their application domains. In this paper, we propose a novel framework of Graph Embedded Contrastive Learning for Multi-View Clustering (GMVC), which unifies MVC and MVGC into a single methodology. GMVC achieves significant clustering performance on both multi-view data and multi-graph data, where the proposed components (i.e., view-specific pre-training, graph-guided contrastive learning, and cluster-guided contrastive clustering) play crucial roles in the multi-view representation learning and clustering processes for both data types. As a result, we aim for our method to enable a unified perspective in addressing the subproblems of MVC and MVGC, while we also hope it can inspire the community to propose more unified methods for both MVC and MVGC.

Acknowledgments

This work was supported in part by the National Key Research & Development Program of China under Grant 2022YFA1004100, the Guangxi Natural Science Foundation (2023GXNSFBA026010), the Sichuan Science and Technology Program (2024NSFSC1473), the Central Guidance for Local Science and Technology Development Fund Project (2024ZYD0268), the Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (ZYGX2022YGRH009 and ZYGX2022YGRH014), the Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education (EBME24-09), the Joint Cultivation Projects of the National Natural Science Foundation of China for Guangxi Normal University in 2024 (2024PY031), the Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security (22-A-03-02).

References

[Abavisani and Patel, 2018] Mahdi Abavisani and Vishal M Patel. Deep multimodal subspace clustering networks. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1601–1614, 2018.

[Bickel and Scheffer, 2004] Steffen Bickel and Tobias Scheffer. Multi-view clustering. In *Proceedings of the IEEE International Conference on Data Mining*, pages 19–26, 2004.

[Chanpuriya and Musco, 2022] Sudhanshu Chanpuriya and Cameron Musco. Simplified graph convolution with heterophily. In *Advances in Neural Information Processing Systems*, pages 27184–27197, 2022.

[Chaudhuri *et al.*, 2009] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *International Conference on Machine Learning*, pages 129–136, 2009.

[Chen *et al.*, 2022] Man-Sheng Chen, Jia-Qi Lin, Xiang-Long Li, Bao-Yu Liu, Chang-Dong Wang, Dong Huang, and Jian-Huang Lai. Representation learning in multi-view clustering: A literature review. *Data Science and Engineering*, 7(3):225–241, 2022.

[Chen *et al.*, 2023] Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16752–16761, 2023.

[Chen *et al.*, 2024] Mulin Chen, Bocheng Wang, and Xuelong Li. Deep contrastive graph learning with clustering-oriented guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11364–11372, 2024.

[Chen *et al.*, 2025] Jianpeng Chen, Yawen Ling, Jie Xu, Yazhou Ren, Shudong Huang, Xiaorong Pu, Zhifeng Hao, Philip S. Yu, and Lifang He. Variational graph generator for multiview graph clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[Cheng *et al.*, 2021] Jiafeng Cheng, Qianqian Wang, Zhiqiang Tao, Deyan Xie, and Quanxue Gao. Multi-view attribute graph convolution networks for clustering. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 2973–2979, 2021.

[Chuang *et al.*, 2022] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16670–16681, 2022.

[Fan *et al.*, 2020] Shaohua Fan, Xiao Wang, Chuan Shi, Emiao Lu, Ken Lin, and Bai Wang. One2multi graph autoencoder for multi-view graph clustering. In *Proceedings of the Web Conference*, pages 3070–3076, 2020.

[Fei-Fei *et al.*, 2004] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004.

[Fu *et al.*, 2025] Yulu Fu, Yuting Li, Qiong Huang, Jinrong Cui, and Jie Wen. Anchor graph network for incomplete multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3708–3719, 2025.

[Hinton and Salakhutdinov, 2006] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[Hussain *et al.*, 2010] Syed Fawad Hussain, Gilles Bisson, and Clément Grimal. An improved co-similarity measure for document clustering. In *International Conference on Machine Learning and Applications*, pages 190–197, 2010.

[Jin *et al.*, 2021] Di Jin, Cuiying Huo, Chundong Liang, and Liang Yang. Heterogeneous graph neural network via attribute completion. In *Proceedings of the Web Conference*, pages 391–400, 2021.

[Jin *et al.*, 2023] Jiaqi Jin, Siwei Wang, Zhibin Dong, Xinwang Liu, and En Zhu. Deep incomplete multi-view clustering with cross-view partial sample and prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11600–11609, 2023.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[Liang *et al.*, 2021] Xinyan Liang, Yuhua Qian, Qian Guo, Honghong Cheng, and Jiye Liang. Af: An association-based fusion method for multi-modal classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9236–9254, 2021.

[Lin and Kang, 2021] Zhiping Lin and Zhao Kang. Graph filter-based multi-view attributed graph clustering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2723–2729, 2021.

[Lin *et al.*, 2012] Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yung-Huan Hsieh, and Hong-Ming Chen. Human action recognition and retrieval using sole depth information. In *Proceedings of the ACM International Conference on Multimedia*, pages 1053–1056, 2012.

[Ling *et al.*, 2023] Yawen Ling, Jianpeng Chen, Yazhou Ren, Xiaorong Pu, Jie Xu, Xiaofeng Zhu, and Lifang He. Dual label-guided graph refinement for multi-view graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8791–8798, 2023.

[Liu *et al.*, 2022] Liang Liu, Zhao Kang, Jiajia Ruan, and Xixu He. Multilayer graph contrastive clustering network. *Information Sciences*, 613:256–267, 2022.

[Luo *et al.*, 2024] Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. Simple contrastive multi-view clustering with data-level fusion. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4697–4705, 2024.

[MacQueen, 1967] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–298, 1967.

[Mo *et al.*, 2024] Yujie Mo, Feiping Nie, Ping Hu, Heng Tao Shen, Zheng Zhang, Xinchao Wang, and Xiaofeng Zhu. Self-supervised heterogeneous graph learning: a homophily and heterogeneity view. In *International Conference on Learning Representations*, 2024.

[Nie *et al.*, 2017] Feiping Nie, Jing Li, and Xuelong Li. Self-weighted multiview clustering with multiple graphs. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2564–2570, 2017.

[Perkins and Yang, 2019] Hugh Perkins and Yi Yang. Dialog intent induction with deep multi-view clustering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 4016–4025, 2019.

[Qian *et al.*, 2024] Xiaowei Qian, Bingheng Li, and Zhao Kang. Upper bounding barlow twins: A novel filter for multi-relational clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14660–14668, 2024.

[Sun *et al.*, 2007] Ting-Kai Sun, Song-Can Chen, Zhong Jin, and Jing-Yu Yang. Kernelized discriminative canonical correlation analysis. In *International Conference on Wavelet Analysis and Pattern Recognition*, pages 1283–1287, 2007.

[Sun *et al.*, 2024] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):9150–9162, 2024.

[Tang and Liu, 2022] Huayi Tang and Yong Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International Conference on Machine Learning*, pages 21090–21110, 2022.

[Trosten *et al.*, 2021] Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1255–1265, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[Wang *et al.*, 2022] Boyue Wang, Yifan Wang, Xiaxia He, Yongli Hu, and Baocai Yin. Multi-graph convolutional clustering network. *IET Signal Processing*, 16(6):650–661, 2022.

[Wen *et al.*, 2021] Jie Wen, Zhihao Wu, Zheng Zhang, Lunke Fei, Bob Zhang, and Yong Xu. Structural deep incomplete multi-view clustering network. In *Proceedings of the ACM International Conference on Information & Knowledge Management*, pages 3538–3542, 2021.

[Xia *et al.*, 2022] Wei Xia, Sen Wang, Ming Yang, Quanxue Gao, Jungong Han, and Xinbo Gao. Multi-view graph embedding clustering network: Joint self-supervision and block diagonal representation. *Neural Networks*, 145:1–9, 2022.

[Xie *et al.*, 2016] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pages 478–487, 2016.

[Xu *et al.*, 2022] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16051–16060, 2022.

[Xu *et al.*, 2023] Jie Xu, Shuo Chen, Yazhou Ren, Xiaoshuang Shi, Hengtao Shen, Gang Niu, and Xiaofeng Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. In *Advances in Neural Information Processing Systems*, pages 1119–1131, 2023.

[Xu *et al.*, 2024] Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, and Xiaofeng Zhu. Investigating and mitigating the side effects of noisy views for self-supervised clustering algorithms in practical multi-view scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22957–22966, 2024.

[Zhou and Shen, 2020] Runwu Zhou and Yi-Dong Shen. End-to-end adversarial-attention network for multi-modal clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14619–14628, 2020.

[Zhou *et al.*, 2003] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, 2003.