Contrastive and View-Interaction Structure Learning for Multi-view Clustering

Jing Wang[1,2] , **Songhe Feng**[1,2*]

[1]Key Laboratory of Big Data & Artificial Intelligence in Transportation (Beijing Jiaotong University), Ministry of Education

[2]School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China

{jing_w, shfeng}@bjtu.edu.cn

Abstract

Existing Deep Multi-view Clustering (DMVC) approaches typically concentrate on capturing consensus semantics from multiple views, where contrastive learning is widely used to align view-specific representations of each view. Unfortunately, view-specific representations are extracted from the content information of the corresponding instance, neglecting the relationships among different instances. Furthermore, existing contrastive loss imports numerous false negative pairs that conflict with the clustering objectives. In response to these challenges, we propose a contraStive and viEw-interaction stRucture learning framework for multI-viEw cluStering (SERIES). Our method takes into account the structural relations among instances and boosts the contrastive loss to improve intra-class compactness. Meanwhile, a cross-view dual relation generation mechanism is introduced to achieve the consensus structural graph across multiple views for clustering. Specifically, we initially acquire view-specific representations using multiple graph autoencoders to exploit both content information and structural information. Furthermore, to pull together the same cluster instances, a soft negative pair aware contrastive loss is employed to distinguish the dissimilar instances while attracting similar instances. Thereafter, the view-specific representations are fed into cross-view dual relation generation layers to generate the affinity matrices of each other, aiming to reveal a consistent structural graph across various views. Extensive experiments conducted on six benchmarks illustrate the superiority of our method compared to other state-of-the-art approaches.

1 Introduction

In contemporary times, data originating from diverse domains, sensors, or feature extractors is readily amassed owing to the prevalence of network edge devices. Recent years have witnessed notable success in Multi-view Clustering (MVC) [Wen *et al.*, 2022], which enhances clustering performance by integrating varied content information from multiple perspectives. MVC [Li *et al.*, 2023; Wen *et al.*, 2023c; Trosten *et al.*, 2023; Wen *et al.*, 2023b] methods can be roughly divided into graph-based [Tan *et al.*, 2023], co-training [Kumar *et al.*, 2011], subspace-based [Tang *et al.*, 2022], multiple kernels-based [Liu *et al.*, 2020], and deep learning-based [Xu *et al.*, 2023; Wen *et al.*, 2023a] approaches respectively. Among the numerous MVC methods, Deep Learning-based Multi-view Clustering (DMVC) stands out for its superior performance attributed to the remarkable representational capacity of deep neural networks. For instance, Multi-VAE [Xu *et al.*, 2021] utilizes a deep generative network to disentangle visual representations into view-common and view-specific features, which are assumed to follow a discrete Gumbel Softmax distribution and a continuous Gaussian distribution, respectively. Completer [Lin *et al.*, 2021] focuses on acquiring complementary multiple representations through a within-view reconstruction task. Simultaneously, it incorporates a contrastive task to leverage the consistency of multiple representations extracted by view-specific deep encoders. Although remarkable progress has been achieved by existing DMVC methods, they merely take advantage of multi-view (content) information from the same sample, neglecting to consider the topological relation (structure) information among different samples. As an unsupervised representation learning task, mining the intrinsic relationship of samples is crucial for multi-view clustering. On the other hand, even contrastive learning significantly improves the performance of unsupervised representation learning, which recognizes all of the other samples as negatives, importing many false negative pairs and increasing the intra-class distance.

In this paper, we propose a contrastive and view-interaction structure learning framework for multi-view clustering (SERIES) to address the aforementioned issues, as shown in Figure 1. SERIES aims to explore the structure relations among different samples and decrease the intra-class distance, finally achieving a consistent structure graph for spectral clustering. Specifically, we first utilize both the multi-view data and structure graphs as the input of view-specific deep graph autoencoders to learn multiple latent representations. Then, the soft negative pair aware contrastive learning module introduces a dynamic sample weighting strategy of negative pairs to pull together similar samples while pushing apart

*Corresponding author

the dissimilar samples in the feature space. According to the view-specific representations, we use a cross-view dual relation generation mechanism to generate the structure graphs for each other, which fully uncovers the consistent topological structure across multiple views. Finally, the view-specific structure graphs are integrated into a unified one for subsequent spectral clustering. The major contributions are summarized as follows:

- We propose a contrastive and view-interaction structure learning framework for multi-view clustering, which is able to reveal the intrinsic structure of data points and achieve a more compact cluster structure. Extensive experiments are conducted to validate the efficacy of our approach.
- Different from the existing DMVC methods, we simultaneously adopt view-specific graph autoencoders and design a cross-view dual relation generation mechanism to capture the intrinsic structure of samples and model the consistent structure graph for clustering.
- To reduce the intra-class distance, we propose a soft negative pair aware contrastive loss, which can utilize a dynamic sample weighting strategy to distinguish similar samples from negative pairs.

2 Related Work

In recent years, advancements in multi-view clustering have demonstrated improved performance through the exploration of complementary information from various views. These approaches can be broadly categorized into co-training, subspace-based, multiple kernels-based, and graph-based methods, respectively. Co-training multi-view clustering [Kumar *et al.*, 2011; Kumar *et al.*, 2011] employs co-regularization on clustering hypotheses to accomplish consistent clustering across all views. Multi-view subspace clustering methods [Huang *et al.*, 2022] usually project multi-view data into a common subspace latent space, uncovering accurate group information for a set of data points. Multiple kernel clustering [Liu *et al.*, 2023] involves computing several kernel matrices for each view, mapping multi-view data into a high-dimensional Hilbert space for clear data point separation. Although these methods exhibit strong performance, they often neglect the structural relations between instances. In contrast, graph-based multi-view clustering methods construct graphs for each view and execute spectral clustering on the consensus graph, allowing for comprehensive capture of topological structure information among instances.

Recently, Graph Neural Network (GNN) [Wu *et al.*, 2020] has drawn lots of attention for its capacity to uncover both structure information and content information within samples. For instance, Graph Convolutional Network (GCN) [Kipf and Welling, 2017] first extends convolutional neural networks to graph structure data, which proposes a layer-wise propagation rule to learn a hidden representation containing both local topological and attributes of nodes. Considering assigning different importance to different nodes and improving efficiency, graph attention networks [Velickovic *et al.*, 2018] introduce masked self-attentional layers into the graph convolutional network to enhance aggregation capability. Owing to its adeptness in leveraging structural relationships, the GNN has been applied to multi-view clustering [Wen *et al.*, 2020a; Wen *et al.*, 2023d]. O2MGC [Fan *et al.*, 2020] is the first attempt to introduce GNN into MVC, which learns the node representation from the structure graph and node content information by GCN encoder and reconstructs multiple graphs using various decoders. After that, SGDMC [Huang *et al.*, 2023] employs an attention-allocating approach to compute node similarity, mitigating the negative impact caused by noisy nodes.

3 Method

In this work, we propose a contrastive and view-interaction structure learning framework for multi-view clustering, which is composed of three submodules and the detailed flowchart is illustrated in Figure 1.

3.1 View-specific Deep Graph Autoencoders

The primary limitation of previous deep multi-view clustering works is that the topological structure is not fully utilized. In this work, we address this gap by employing graph neural networks to propagate structure relations during the representation learning stage.

Given the multi-view data $\{\mathbf{X}^v \in \mathbb{R}^{n\times d_v}\}_{v=1}^m$ encompasses n samples and v views, where d_v represents the feature dimension of the v-th view, we explicitly construct view-specific graphs $\{\mathbf{A}^v \in \mathbb{R}^{n\times n}\}_{v=1}^m$ using the k-Nearest Neighbors (kNN) algorithm. Then, in order to learn the multiple representations containing both content and structure information, we introduce the graph autoencoder, which consists of view-specific graph encoders and view-specific graph decoders:

View-specific graph encoders. The graph encoders can enhance view-specific representations by aggregating neighbor information through structure graphs. Specifically, for v-th view, we feed both feature matrix $\mathbf{X}^v$ and graph $\mathbf{A}^v$ into a multi-layer GCN to learn the instance-level latent representation $\mathbf{Z}^v$, and the computation of the l-th layer of GCN is formulated as follows:

$$\begin{aligned}\mathbf{Z}^{(v,l)} &= f^{(v,l)}(\mathbf{Z}^{(v,l-1)}, \hat{\mathbf{A}}^v; \theta^v) \\ &= \phi(\widetilde{\mathbf{D}}_v^{\frac{1}{2}}\widetilde{\mathbf{A}}^v\widetilde{\mathbf{D}}_v^{\frac{1}{2}}\mathbf{Z}^{(v,l-1)}\mathbf{W}^{(v,l)} + \mathbf{b}^{(v,l)})\end{aligned} \tag{1}$$

where $\widetilde{\mathbf{A}}^v = \mathbf{A}^v + \mathbf{I}_n \in \mathbb{R}^{n\times n}$ is the affinity matrix with added self-connections, and $\widetilde{\mathbf{D}}_{ii}^v = \sum_j \mathbf{A}_{ij}^v \in \mathbb{R}^{n\times n}$ denotes the corresponding degree matrix. $\phi(\cdot)$ is the activation function, $\mathbf{W}^{(v,l)}$ and $\mathbf{b}^{(v,l)}$ serve as parameters for the l-th GCN layer. When $l = 0$, $\mathbf{Z}^{(v,0)}$ corresponds to the original data $\mathbf{X}^v$ of the v-th view, and the latent representation $\mathbf{Z}^{(v,1)}$ is obtained by:

$$\mathbf{Z}^{(v,1)} = \phi(\widetilde{\mathbf{D}}_v^{\frac{1}{2}}\widetilde{\mathbf{A}}^v\widetilde{\mathbf{D}}_v^{\frac{1}{2}}\mathbf{X}^v\mathbf{W}^{(v,1)} + \mathbf{b}^{(v,1)}) \tag{2}$$

View-specific graph decoders. In order to guide the view-specific representations to maintain the instance structure relationship in the latent space, we employ view-specific graph

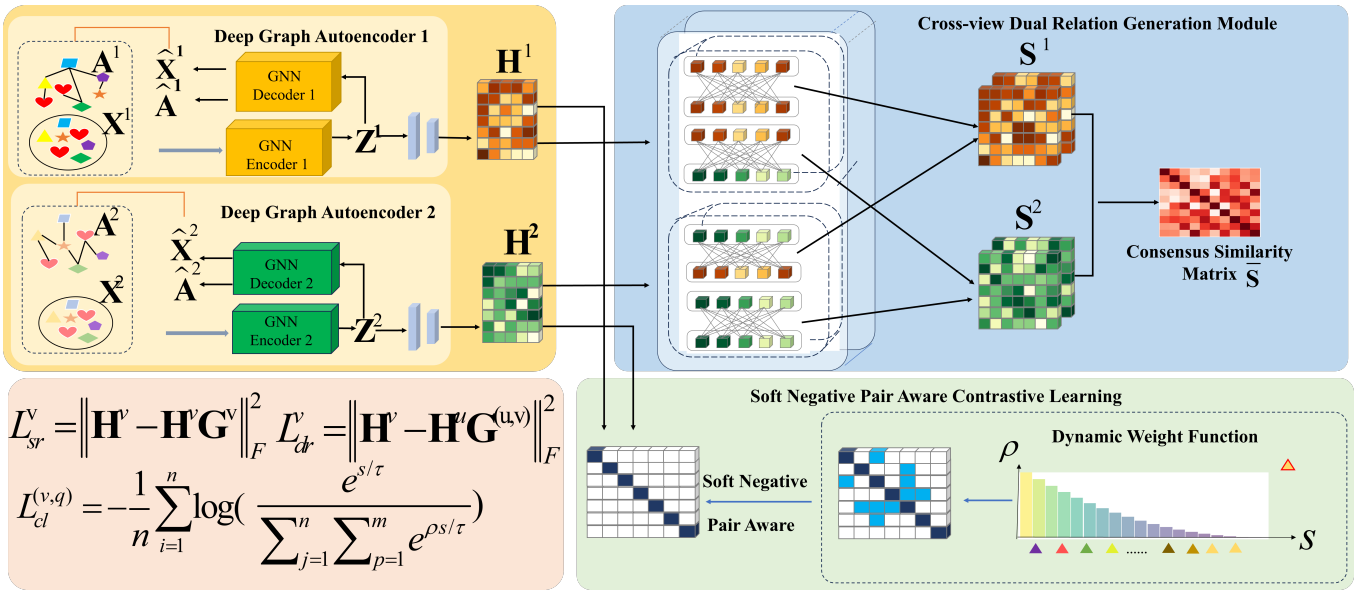


Figure 1: Overview of SERIES. In this figure, we use two view data as a showcase. As described in the figure, our model is composed of three submodules: view-specific deep graph autoencoders, soft negative pair aware contrastive learning module, and cross-view dual relation generation module. Concretely, the view-specific deep graph autoencoders are adopted to extract the latent representations $\{\mathbf{Z}^v\}_{v=1}^m$ that contains both the content and graph information. The soft negative pair aware contrastive learning module is implemented to better pull similar samples closer and push dissimilar samples farther away in the feature space. The cross-view dual relation generation module utilizes both own and other views to generate the view-specific affinity matrix, exploring the consistent structure graph $\overline{\mathbf{S}}$. Finally, $\overline{\mathbf{S}}$ is treated as the input of the spectral clustering method to get clustering results.

decoders to reconstruct graph $\hat{\mathbf{A}}^v$. Specifically, the inner product decoder is adopted to predict the links between instance i and j, which can be formulated as:

$$\hat{\mathbf{A}}^v = sigmoid(\mathbf{Z}^v\mathbf{W}^v\mathbf{Z}^{v^T}) \tag{3}$$

The graph reconstruction error measuring the difference between $\hat{\mathbf{A}}^v$ and $\mathbf{A}^v$ of all views can be calculated by:

$$\mathcal{L}_{g\_rec} = \frac{1}{m}\sum_{v=1}^{m}\left\|\mathbf{A}^v - \hat{\mathbf{A}}^v\right\|_F^2 \tag{4}$$

Meanwhile, to embed the content information of multi-view data into instance latent representation $\mathbf{Z}^v$, the content decoders are introduced to reconstruct $\hat{\mathbf{X}}^v$. More specifically, for the v-th view, a view-specific decoder network $g^v(\cdot)$, parametrized by μ^v, is employed to decode $\mathbf{Z}^v$ into $\hat{\mathbf{X}}^v$:

$$\hat{\mathbf{X}}^v = g^v(\mathbf{Z}^v;\mu^v) \tag{5}$$

The content reconstruction loss of all views is defined as follows:

$$\mathcal{L}_{c\_rec} = \frac{1}{m}\sum_{v=1}^{m}\left\|\mathbf{X}^v - \hat{\mathbf{X}}^v\right\|_F^2 \tag{6}$$

3.2 Soft Negative Pair Aware Contrastive Learning Module

As a key module of deep learning-based multi-view clustering methods, contrastive learning is adopted to align multiple representations from different views of each instance and learn the corresponding discriminative features. Specifically, for mn samples $\{\mathbf{X}_1^1,\ldots,\mathbf{X}_i^1,\ldots,\mathbf{X}_i^m,\ldots,\mathbf{X}_n^m\}$, standard contrastive learning typically treats $(\mathbf{X}_i^v,\mathbf{X}_i^u)$ as positive pair and other $mn-m$ samples to be negative pairs. However, samples that belong to the same cluster should not be treated as negative pairs, which conflicts with the purpose of the clustering task. To address the problem, many works [Trosten *et al.*, 2021; Xia *et al.*, 2023] introduce the pseudo labels to remove within-class samples from negative pairs. Although this adjustment mitigates the issue of false negative pairs, it is difficult to achieve reliable pseudo-labels during the training process, especially in the initial stage of training.

Different from previous works that categorize negative pairs by hard pseudo-labels, we propose a soft negative pair aware contrastive loss, aiming to bring correlated samples closer while distinguishing uncorrelated ones. Specifically, we introduce a weight modulating function $\rho(\cdot,\cdot)$ to dynamically adjust the weights of the sample pairs during training. Based on the view-specific latent representation $\mathbf{Z}_i^v$, $\rho(\cdot,\cdot)$ can be defined as follows:

$$\rho(\mathbf{h}_i^v,\mathbf{h}_j^p) = (1 - s(\mathbf{h}_i^v,\mathbf{h}_j^p))^\beta \tag{7}$$

where $\mathbf{h}_i^v$ is obtained by a view-shared project head $\mathbf{h}_i^v = \sigma(\mathbf{z}_i^{(v,l)})$, and $\sigma(\cdot)$ is introduced to filter out the view-specific noise. $s(\cdot,\cdot)$ is the similarity function, which can be computed as $s(\mathbf{h}_i^v,\mathbf{h}_j^p) = (\mathbf{h}_i^v)^T\mathbf{h}_j^p$. $\beta \in [1,5]$ is the penalty factor employed to adjust the degree of penalization for both uncorrelated samples and correlated samples. For instance, when $\beta = 2$, the similarity of the correlated samples pair

is 0.8, but the corresponding weight is 0.04. In contrast, for uncorrelated samples pair, the similarity is 0.1, while the resulting weight is 0.81, which significantly surpasses 0.04.

According to the dynamic weight modulating function $\rho(\cdot,\cdot)$, we formulate the soft negative pair aware contrastive loss between view v and u as follows:

$$\mathcal{L}_{cl}^{(v,u)} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s(\mathbf{h}_i^v,\mathbf{h}_i^u)/\tau}}{\sum_{j=1}^{n}\sum_{p=1}^{m}e^{\rho(\mathbf{h}_i^v,\mathbf{h}_j^u))s(\mathbf{h}_i^v,\mathbf{h}_j^p)/\tau}} \quad (8)$$

where τ denotes the temperature parameter. The soft negative pair aware contrastive loss across all views is defined as:

$$\mathcal{L}_{cl} = \frac{1}{2}\sum_{v=1}^{m}\sum_{v\neq u}\mathcal{L}_{cl}^{(v,u)} \quad (9)$$

3.3 Cross-view Dual Relation Generation Module

Through the graph autoencoders and the soft negative pair-aware contrastive learning module, we can obtain discriminative feature $\mathbf{H}^v$, which contains both the structure and content information. To fully fuse complementary information across multiple views, we introduce a cross-view dual relation generation module in this subsection. This module generates view-specific affinity matrices by leveraging information from both the target view and other views, facilitating the exploration of a consistent topological structure graph.

Specifically, we first employ the self-relation generation layer $\mathbf{SR}(\cdot)$ to generate the affinity matrix $\mathbf{G}^v$ of v-th view, which aims to represent each instance as a combination of others from v-th view. The generation proposes is formulated as follows:

$$\mathbf{G}^v = \mathbf{SR}(\mathbf{H}^v, \mathbf{W}_s) \quad (10)$$

To explore the global structure of samples under the current view, we minimize the following reconstruction loss:

$$\mathcal{L}_{sr}^v = \|\mathbf{H}^v - \mathbf{H}^v\mathbf{G}^v\|_F^2 \quad (11)$$

Furthermore, to fully integrate the complementary information across different views, we propose a dual-relation generation layer $\mathbf{DR}^u(\cdot)$ to generate the affinity matrix $\mathbf{G}^{(u,v)}$ of v-th view, which can be formulated as:

$$\mathbf{G}^{(u,v)} = \mathbf{DR}^u(\mathbf{H}^u, \mathbf{W}_d) \quad (12)$$

where $\mathbf{G}^{(u,v)}$ is the affinity matrix generated by u-th view. It utilizes the representations $\mathbf{H}^u$ from u-th views to represent $\mathbf{H}^v$, and the reconstruction loss is formulated as follows:

$$\mathcal{L}_{dr}^v = \|\mathbf{H}^v - \mathbf{H}^u\mathbf{G}^{(u,v)}\|_F^2 \quad (13)$$

The overall reconstruction loss can be computed as follows:

$$\mathcal{L}_{str}^v = \mathcal{L}_{sr}^v + \frac{1}{m-1}\sum_{u=1,u\neq v}^{m}\mathcal{L}_{dr}^u \quad (14)$$

Accordingly, the affinity matrix of v-th view is obtained by:

$$\mathbf{S}^v = \frac{1}{2}(\mathbf{G}^v + \frac{1}{m-1}\sum_{u=1,u\neq v}^{m}\mathbf{G}^{(u,v)}) \quad (15)$$

In this way, the view-specific affinity matrix $\mathbf{S}^v$ has the ability to fuse the complementary structure relations from multiple views, and the consensus affinity matrix $\overline{\mathbf{S}}$ is simply obtained by:

$$\overline{\mathbf{S}} = \frac{1}{m}\sum_{v=1}^{m}\mathbf{S}^{(v)} \quad (16)$$

3.4 The Overall Loss Function of SERIES

In summary, we have introduced a contrastive and view-interaction structure learning framework for multi-view clustering. In the training stage, the view-specific graph autoencoders, soft negative pair aware contrastive learning module, and the dual relation generation module are jointly optimized according to the following objective function:

$$\mathcal{L} = \frac{1}{m}\sum_{v=1}^{m}(\mathcal{L}_{g\_rec}^v + \mathcal{L}_{c\_rec}^v + \lambda_1\mathcal{L}_{str}^v) + \lambda_2\mathcal{L}_{cl} \quad (17)$$

Finally, we obtain a desirable consensus affinity matrix $\overline{\mathbf{S}}$ and pass it through a spectral clustering algorithm to achieve the final clustering result. The whole learning process of SERIES is summarized in the Algorithm 1.

Algorithm 1 The Algorithm of **SERIES**

Input: Multi-view data $\{\mathbf{X}^{(v)}\}_{v=1}^m$; Training iterations T.
Process:
1. Construct the graphs for each view and obtain the view-specific affinity matrices $\{\mathbf{A}^{(v)}\}_{v=1}^m$.
Pretrain:
2. Pertrain the deep graph autoencoders of each view by optimizing $\mathcal{L}_{g\_rec}^v$, $\mathcal{L}_{c\_rec}^v$ in Eq. 4 and 6.
Finetuning:
3. for epoch = 1 to T
4. Obtain the $\mathbf{Z}_v, \mathbf{G}_v, \mathbf{G}^{(u,v)}, \mathbf{S}_v$ of each view by Eqs. (1,10,12,15).
5. Update network parameters by using Adam to minimize the objective in Eq. 17.
9. end for
return: The consensus affinity matrix $\overline{\mathbf{S}}$.
Perform spectral clustering using $\overline{\mathbf{S}}$.

4 Experiment

4.1 Experimental Settings

Datasets. The following datasets are carried out for evaluation: (1) **HW** [Perkins and Theiler, 2003] consists of 2000 samples from 10 types of handwritten digits, all of which are presented by two features. (2) **Reuters** [Amini *et al.*, 2009] comprises 6 categories, corresponding to 1200 articles, and each article is written in 5 languages. (3) **Noisyminist** [Wang *et al.*, 2015] contains 70k instances described by

Datasets	HW						Mfeat					
Method	**ACC**	**NMI**	**PUR**	**ARI**	**P**	**F-score**	**ACC**	**NMI**	**PUR**	**ARI**	**P**	**F-score**
LMSC	0.7720	0.6504	0.7720	0.5931	0.6292	0.6339	0.7235	0.6190	0.7235	0.5326	0.5772	0.5793
MCGC	0.8775	0.8820	0.8780	0.8480	0.7925	0.8645	0.9540	0.9070	0.9540	0.9005	0.9083	0.9105
LMVSC	0.8195	0.7665	0.8195	0.6769	0.7337	0.7101	0.6550	0.6386	0.7461	0.5283	0.6218	0.5788
CDIMIC	0.8285	0.8982	0.7780	0.8199	0.8569	0.8793	0.8360	0.8882	0.8705	0.8145	0.8423	0.8588
SiMVC	0.8401	0.8448	0.7827	0.7959	0.8223	0.7757	0.8001	0.8407	0.8413	0.7563	0.7853	0.8047
CoMVC	0.9128	0.9042	0.8836	0.8862	0.9070	0.8689	0.7750	0.8243	0.8147	0.7207	0.761	0.7757
MFLVC	0.7830	0.7826	0.7830	0.6780	0.7174	0.7202	0.8300	0.8093	0.830	0.7358	0.7641	0.7674
CMGEC	0.7250	0.7009	0.7295	0.5467	0.7670	0.7290	0.7170	0.7388	0.7380	0.6341	0.6977	0.6951
DFP-GNN	0.9505	0.9024	0.9515	0.8936	0.9074	0.9076	0.9490	0.9070	0.8520	0.8900	0.9088	0.9082
SERIES	**0.9665**	**0.9288**	**0.9665**	**0.9663**	**0.9367**	**0.9360**	**0.9655**	**0.9252**	**0.9665**	**0.9240**	**0.9348**	**0.9340**
Datasets	**Noisyminist**						**VOC**					
Method	**ACC**	**NMI**	**PUR**	**ARI**	**P**	**F-score**	**ACC**	**NMI**	**PUR**	**ARI**	**P**	**F-score**
LMSC	0.2734	0.2039	0.3038	0.1285	0.2113	0.2279	0.1837	0.1246	0.2822	0.0657	0.1693	0.1252
MCGC	0.4863	**0.5399**	0.4971	**0.4063**	0.3390	**0.4930**	0.2935	0.1387	0.2953	0.1116	0.1395	0.2337
LMVSC	0.3274	0.3027	0.5196	0.1603	0.3858	0.2699	0.1637	0.1357	0.1772	0.0523	0.0939	0.1142
CDIMIC	0.4812	0.4527	0.4857	0.3490	0.3858	0.4343	0.1855	0.1346	0.2857	0.0520	0.1519	0.1451
SiMVC	0.3831	0.3266	0.4109	0.2988	0.2923	0.2163	0.5376	0.5511	0.6640	0.4788	0.5533	0.4806
CoMVC	0.4141	0.4047	0.4667	0.3616	0.3469	0.2674	0.5151	0.5307	0.6435	0.4173	0.5358	0.4579
MFLVC	0.2497	0.2054	0.1905	0.0778	0.1905	0.2609	0.5249	0.4570	0.5304	0.3152	0.3566	0.4433
CMGEC	OM	OM	OM	OM	OM	OM	0.3234	0.3397	0.3987	0.1643	0.3004	0.1826
DFP-GNN	0.4649	0.4416	0.5566	0.2864	0.4051	0.3758	0.6113	0.5350	0.6375	0.4731	0.4998	0.4925
SERIES	**0.5189**	0.5275	**0.5683**	0.3896	**0.4730**	0.4789	**0.7325**	**0.6867**	**0.7743**	**0.5516**	**0.7011**	**0.6521**
Datasets	**Hdigit**						**Reuters**					
Method	**ACC**	**NMI**	**PUR**	**ARI**	**P**	**F-score**	**ACC**	**NMI**	**PUR**	**ARI**	**P**	**F-score**
LMSC	0.6681	0.6207	0.7151	0.5269	0.5625	0.5753	0.4450	0.2635	0.4783	0.1985	0.3134	0.3449
MCGC	0.5814	0.6339	0.5816	0.5386	0.4488	0.6002	0.1850	0.0426	0.2075	0.0032	0.1673	0.2836
LMVSC	0.5482	0.5050	0.6045	0.3544	0.4937	0.4275	0.3692	0.1920	0.6192	0.1196	**0.4907**	0.3173
CDIMIC	0.5071	0.5412	0.5199	0.3584	0.4361	0.4820	0.1842	0.0565	0.3483	0.0030	0.1916	0.3182
SiMVC	0.7435	0.7638	0.7564	0.6893	0.6975	0.6541	0.2895	0.0615	0.3605	0.2060	0.2165	0.0474
CoMVC	0.8027	0.8244	0.8175	0.7680	0.7759	0.7349	0.2940	0.0659	0.3703	0.2087	0.2138	0.0488
MFLVC	0.9478	0.8834	0.9478	0.8885	0.9002	0.9003	0.4550	0.2371	0.455	0.1853	0.3236	0.3238
CMGEC	0.3149	0.1843	0.3564	0.1100	0.3283	0.3186	0.2200	0.0229	0.3333	0.0082	0.2359	0.2183
DFP-GNN	0.8847	0.8810	0.8919	0.8312	0.8414	0.8537	0.6042	0.3961	**0.6558**	0.2954	0.4557	0.4571
SERIES	**0.9676**	**0.9232**	**0.9679**	**0.9300**	**0.9372**	**0.9372**	**0.6358**	**0.4134**	0.6383	**0.3131**	0.4800	**0.4840**

Table 1: The clustering performance comparisons on six multi-view datasets.

2 views. In this study, we choose a subset of Noisyminist, consisting of 15,000 instances from 10 classes, for the comparative experiment. (4) **VOC** [Hwang and Grauman, 2010] is a two-view dataset that encompasses image-view and text-view modalities, which consists of 5,649 instances distributed across 20 distinct classes. (5) **Hdigit** [Chen *et al.*, 2022] includes 10,000 instances categorized into 10 classes, whose two views are crafted from both MNIST Handwritten Digits and USPS Handwritten Digits. (6) **Mfeat** [Wang *et al.*, 2019] comprises 2000 instances from 10 subjects, with six features extracted to form this multi-view dataset. The detailed information is summarized in Table 2.

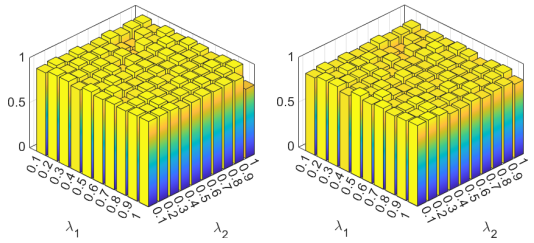
Metrics. We employ six widely used metrics to evaluate our model, including Accuracy (ACC), Normalized Mutual Information (NMI), Purity (PUR), Adjusted Rand Index (ARI), Precision (P), and F-score. The specific definitions of these metrics are explained in [Cao *et al.*, 2015].

Baselines. Our proposed method is compared with nine state-of-the-art methods, which are summarized as follows:

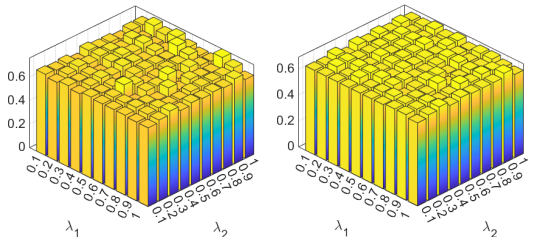
- **LMSC** [Zhang *et al.*, 2017] utilizes a self-supervised reconstruction task to acquire latent representations for multiple views while concurrently investigating the inherent complementarity for MVC.
- **MCGC** [Zhan *et al.*, 2019] dynamically learns a consensus graph to reveal more robust relationship between data points, which has exactly k connected components aligning with the number of clusters.
- **LMVSC** [Kang *et al.*, 2020] finds some representative data points as anchors to construct anchor graphs, which can present the global structure relations and are beneficial to alleviate the computational issue.
- **CDIMIC** [Wen *et al.*, 2020b] introduces the deep autoencoders and self-paced strategy to extract the high-level features and reduce the negative impact of outliers.
- **SiMVC** [Trosten *et al.*, 2021] is a simple deep MVC model obtaining a fused representation by weighted average multiples representations, which is proposed to prioritize views in the feature space.
- **CoMVC** [Trosten *et al.*, 2021] extends SiMVC by incorporating a contrastive alignment mechanism, which aims at increasing the distance between distinct clusters.

- **MFLVC** [Xu *et al.*, 2022] utilizes contrastive regularization terms to align the instance-level and cluster-level representations, which aims to explore multi-view consistency.
- **CMGEC** [Wang *et al.*, 2021] is a graph-based deep multi-view clustering method, which considers the structure information among samples and introduces a mutual information maximization module to discover the consistent topological structure.
- **DFP-GNN** [Xiao *et al.*, 2023] adopts GNN models and a new dual fusion mechanism to learn a unified representation, which combines both content and structure information across different views.

Implementations. All experiments are conducted on a Linux platform utilizing an Intel(R) Core(TM) i9-11900 2.50GHz CPU, 64GB RAM, and GeForce RTX 3090 Ti GPU. The view-specific deep graph autoencoders are pre-trained for 200 epochs, and the entire model is fine-tuned for an additional 100 epochs. The dimensions of the encoders, decoders, and the cross dual relation generation layer are set to $\{d_v, 512, 2048, 256\}$, $\{256, 2048, 512, d_v\}$ and $\{256, d_v\}$ respectively. The activation function is specified as ReLU. In our study, the trade-off hyperparameters λ_1, λ_2 are selected from the range $\{0.1, 0.2, \ldots, 0.9, 1.0\}$. All other baselines are implemented by the recommended network structure and parameters for fair comparison.



(a) Mfeat(ACC) (b) Mfeat(NMI)



(c) VOC(ACC) (d) VOC(NMI)

Figure 2: Parameters sensitivity analysis: the clustering performances (ACC, NMI) with different parameters λ_1 and λ_2 on VOC and Mfeat datasets

4.2 Comparison Results

Table 1 reports the clustering results of our method and other baselines on six datasets, where the best and the sub-optimal performance are denoted in bold and underlined, respectively. **OM** represent "Out-of-memory error". From the results, we have the following observations:

Datasets	Samples	Clusters	Views	View dimensions
HW	2000	10	6	216/76/64/6/240/47
Reuters	1200	6	5	2000/2000/2000/2000/2000
Noisyminist	15000	10	2	784/784
VOC	5649	20	2	512/399
Hdigit	10000	10	2	784/256
Mfeat	2000	10	6	216/76/64/6/240/47

Table 2: Statistical characteristics of six datasets.

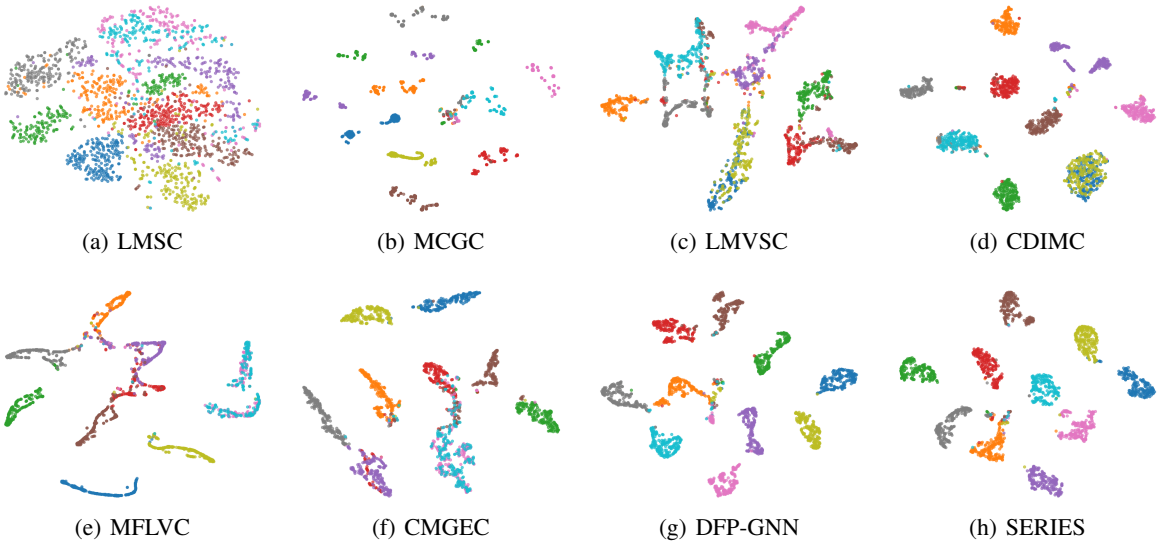
Datasets	Method	ACC	NMI	PUR	ARI
Mfeat	**SERIES**	**0.9650**	**0.9252**	**0.9665**	**0.9240**
	SERIES-D	0.8385	0.8606	0.8655	0.7836
	SERIES-SC	0.9090	0.8434	0.9100	0.8157
	SERIES-C	0.8958	0.8565	0.8933	0.9074
VOC	**SERIES**	**0.7325**	**0.6867**	**0.7743**	**0.5516**
	SERIES-D	0.7113	0.6761	0.7582	0.4756
	SERIES-SC	0.7242	0.6807	0.7654	0.5323
	SERIES-C	0.7247	0.6813	0.7663	0.5342

Table 3: Ablation study on VOC and Mfeat dataset.

(1) Among all the compared methods, SERIES achieves superior performance in most cases, and improvement of our algorithm on some datasets is significant. For example, on the VOC dataset, our algorithm surpasses the sub-optimal algorithm SiMVC by 36.25%, 24.60%, 16.61%, 15.20%, 26.71%, 35.68% in terms of six metrics, respectively. For Hdigit dataset, our algorithm improves 2.08%, 4.50%, 2.12%, 4.67%, 4.11%, 4.13% across six metrics compared to the second-best DFP-GNN method. These observations validate the excellent effectiveness of our model, which deeply investigates the topology structure between samples benefiting the partition of similar samples into the same cluster.

(2) Compared with other state-of-the-art contrastive learning-based multi-view clustering methods, i.e., CoMVC and MFLVC, our method still shows significant advantages. MFLVC treats all of the other samples as negative samples which inevitably brings in some false negative pairs. CoMVC introduces the pseudo-label information to guide the selection of positive and negative pairs, which heavily depends on the quality of pseudo-labels. However, the contrastive loss proposed in our method aims to mitigate the impact of false negative pairs by introducing a dynamic weighting strategy, which reduces the distance between relevant samples and increases the distance between irrelevant samples.

(3) Observed from Table 1, the shallow methods are inferior to the deep learning-based approaches in most cases, particularly on the VOC and Hdigit datasets. This discrepancy can be attributed to the deep learning-based multi-view methods effectively capturing the intrinsic features of instances. Notably, the graph-based method DFP-GNN, surpasses some deep learning-based methods, indicating the significance of exploring structural information between samples. Nevertheless, our approach consistently outperforms both other graph-based and deep learning-based multi-view clustering methods across all datasets. This consistency highlights the efficacy of our cross-view dual relation generation module and soft negative pair aware contrastive learning module.



(a) LMSC (b) MCGC (c) LMVSC (d) CDIMC

(e) MFLVC (f) CMGEC (g) DFP-GNN (h) SERIES

Figure 3: The visualizations of the consensus representation of LMSC, MCGC, LMVSC, CDIMIC, MFLVC, CMGCE, DFP-GNN, and SERIES on Mfeat dataset.

4.3 Ablation Study

To verify the effectiveness of each submodule, we compare our proposed SERIES with its degenerated methods to conduct the ablation study in this subsection. Concretely, in Table 3 we analyze the following cases:

- **SERIES-D**: The affinity matrix $\mathbf{S}^v$ is only composed of $\mathbf{G}^v$, omitting the dual generation process in Eq. 12.
- **SERIES-SC**: SERIES w.o. the soft negative pair aware contrastive loss in Eq. 8.
- **SERIES-C**: The soft negative pair aware contrastive loss is replaced by the standard contrastive loss [Chen *et al.*, 2020], which treats all of the other samples as negative pairs.

From the results in Table 3, we have two observations as follows: (1) In comparison to SERIES-SC and SERIES-D, SERIES exhibits the best performance, indicating that both the soft negative pair aware contrastive learning module and cross-view dual relation generation module enhance the performance of the baselines. These modules facilitate the exploration of structural relations between instances and the acquisition of instance intrinsic features. (2) The results of SERIES-C are consistently lower than those of SERIES-SC across all metrics on both Mfeat and VOC datasets. This observation indicates that our proposed soft negative pair aware contrastive loss effectively mitigates the issue of false negative pairs. Unlike the standard contrastive loss which aims to increase the distance between one sample and all other samples, our method focuses on dynamically adjusting the weight according to the similarity of the samples, which is more suitable for the final clustering task.

4.4 Parameter Sensitivity Analysis

In this subsection, we analyze the hyper-parameters λ_1 and λ_2 in our method on Mfeat and VOC datasets. Figure 2 illustrates the ACC and NMI of our approach as λ_1 and λ_2 vary within the range of $\{0.1, \cdots, 1.0\}$. As depicted in Figure 2, our method consistently demonstrates strong performance across a broad range, particularly exhibiting promising results when $\lambda_1 \in [0.1, 0.5]$ and $\lambda_2 \in [0.5, 1.0]$ on Mfeat and VOC datasets.

4.5 Visualization Analysis

In this subsection, we utilize t-SNE [Van der Maaten and Hinton, 2008] to visualize the final consensus representations in our method and other compare methods. As shown in Figure 3, the consensus representation learned in our method is more compact than others, which demonstrates our proposed SERIES could better explore the cluster structure compared with other baselines.

5 Conclusions

In this work, we propose a contrastive and view-interaction structure learning framework for multi-view clustering, named SERIES, to better reveal the structure relation between different instances. We first utilize the view-specific graph autoencoders to achieve latent representations containing both content and graph information. Then, the soft negative pair aware contrastive learning module introduces a dynamic sample weighting strategy to alleviate the false negative pair problem, which could better learn discriminative features for instances. Finally, to cluster samples with the structure graph of samples, we adopt the cross-view dual relation generation module to explore the consistent affinity matrix across multiple views. The effectiveness of SERIES has been verified on various multi-view datasets as compared with other state-of-the-art methods.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (No.2022JBZY019), the Beijing Natural Science Foundation (No. 4242046).

References

[Amini *et al.*, 2009] Massih-Reza Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views - an application to multilingual text categorization. In *Advances in Neural Information Processing Systems*, pages 28–36, 2009.

[Cao *et al.*, 2015] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–594, 2015.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020.

[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*, pages 225–241, 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.

[Huang *et al.*, 2022] Shudong Huang, Yixi Liu, Ivor W Tsang, Zenglin Xu, and Jiancheng Lv. Multi-view subspace clustering by joint measuring of consistency and diversity. *IEEE Transactions on Knowledge and Data Engineering*, pages 8270–8281, 2022.

[Huang *et al.*, 2023] Zongmo Huang, Yazhou Ren, Xiaorong Pu, Shudong Huang, Zenglin Xu, and Lifang He. Self-supervised graph attention networks for deep weighted multi-view clustering. In *Conference on Artificial Intelligence*, pages 7936–7943, 2023.

[Hwang and Grauman, 2010] Sung Ju Hwang and Kristen Grauman. Accounting for the relative importance of objects in image retrieval. In *BMVC*, page 5, 2010.

[Kang *et al.*, 2020] Zhao Kang, Wangtao Zhou, Zhitong Zhao, Junming Shao, Meng Han, and Zenglin Xu. Large-scale multi-view subspace clustering in linear time. In *Conference on Artificial Intelligence*, pages 4412–4419, 2020.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, pages 1–14, 2017.

[Kumar *et al.*, 2011] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. *Advances in Neural Information Processing Systems*, pages 1–9, 2011.

[Li *et al.*, 2023] Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. Incomplete multi-view clustering via prototype-based imputation. *arXiv preprint*, pages 1–9, 2023.

[Lin *et al.*, 2021] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11174–11183, 2021.

[Liu *et al.*, 2020] Xinwang Liu, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu. Efficient and effective regularized incomplete multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2634–2646, 2020.

[Liu *et al.*, 2023] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Qing Liao, and Yuanqing Xia. Contrastive multi-view kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 9552–9566, 2023.

[Perkins and Theiler, 2003] Simon Perkins and James Theiler. Online feature selection using grafting. In *International Conference on Machine Learning*, pages 592–599, 2003.

[Tan *et al.*, 2023] Yuze Tan, Yixi Liu, Hongjie Wu, Jiancheng Lv, and Shudong Huang. Metric multi-view graph clustering. In *Conference on Artificial Intelligence*, pages 9962–9970, 2023.

[Tang *et al.*, 2022] Kewei Tang, Kaiqiang Xu, Wei Jiang, Zhixun Su, Xiyan Sun, and XiaoNan Luo. Selecting the best part from multiple laplacian autoencoders for multi-view subspace clustering. *IEEE Transactions on Knowledge and Data Engineering*, pages 7457–7469, 2022.

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

[Trosten *et al.*, 2023] Daniel J Trosten, Sigurd Løkse, Robert Jenssen, and Michael C Kampffmeyer. On the effects of self-supervision and contrastive alignment in deep multi-view clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 23976–23985, 2023.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning research*, pages 2579–2605, 2008.

[Velickovic *et al.*, 2018] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, and Adriana Romero. Graph attention networks. In *International Conference on Learning Representations*, pages 1–12, 2018.

[Wang *et al.*, 2015] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015.

[Wang *et al.*, 2019] Hao Wang, Yan Yang, and Bing Liu. Gmc: Graph-based multi-view clustering. *IEEE Transac-*

tions on Knowledge and Data Engineering, pages 1116–1129, 2019.

[Wang *et al.*, 2021] Yiming Wang, Dongxia Chang, Zhiqiang Fu, and Yao Zhao. Consistent multiple graph embedding for multi-view clustering. *IEEE Transactions on Multimedia*, pages 1008–1018, 2021.

[Wen *et al.*, 2020a] Jie Wen, Ke Yan, Zheng Zhang, Yong Xu, Junqian Wang, Lunke Fei, and Bob Zhang. Adaptive graph completion based incomplete multi-view clustering. *IEEE Transactions on Multimedia*, pages 2493–2504, 2020.

[Wen *et al.*, 2020b] Jie Wen, Zheng Zhang, Yong Xu, Bob Zhang, Lunke Fei, and Guo-Sen Xie. Cdimc-net: Cognitive deep incomplete multi-view clustering network. In *International Joint Conference on Artificial Intelligence*, pages 3538–3542, 2020.

[Wen *et al.*, 2022] Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang, and Jinxing Li. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pages 1136–1149, 2022.

[Wen *et al.*, 2023a] Jie Wen, Chengliang Liu, Shijie Deng, Yicheng Liu, Lunke Fei, Ke Yan, and Yong Xu. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2023.

[Wen *et al.*, 2023b] Jie Wen, Chengliang Liu, Gehui Xu, Zhihao Wu, Chao Huang, Lunke Fei, and Yong Xu. Highly confident local structure based consensus graph learning for incomplete multi-view clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15712–15721, 2023.

[Wen *et al.*, 2023c] Jie Wen, Gehui Xu, Chengliang Liu, Lunke Fei, Chao Huang, Wei Wang, and Yong Xu. Localized and balanced efficient incomplete multi-view clustering. In *ACM International Conference on Multimedia*, pages 2927–2935, 2023.

[Wen *et al.*, 2023d] Jie Wen, Gehui Xu, Zhanyan Tang, Wei Wang, Lunke Fei, and Yong Xu. Graph regularized and feature aware matrix factorization for robust incomplete multi-view clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–14, 2023.

[Wu *et al.*, 2020] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 4–24, 2020.

[Xia *et al.*, 2023] Wei Xia, Tianxiu Wang, Quanxue Gao, Ming Yang, and Xinbo Gao. Graph embedding contrastive multi-modal representation learning for clustering. *IEEE Transactions on Image Processing*, pages 1170–1183, 2023.

[Xiao *et al.*, 2023] Shunxin Xiao, Shide Du, Zhaoliang Chen, Yunhe Zhang, and Shiping Wang. Dual fusion-propagation graph neural network for multi-view clustering. *IEEE Transactions on Multimedia*, pages 1–13, 2023.

[Xu *et al.*, 2021] Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9234–9243, 2021.

[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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16051–16060, 2022.

[Xu *et al.*, 2023] Jie Xu, Shuo Chen, Yazhou Ren, Xiaoshuang Shi, Heng Tao Shen, Gang Niu, and Xiaofeng Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. In *Conference on Neural Information Processing Systems*, pages 1–13, 2023.

[Zhan *et al.*, 2019] Kun Zhan, Feiping Nie, Jing Wang, and Yi Yang. Multiview consensus graph clustering. *IEEE Transactions on Image Processing*, pages 1261–1270, 2019.

[Zhang *et al.*, 2017] Changqing Zhang, Qinghua Hu, Huazhu Fu, Pengfei Zhu, and Xiaochun Cao. Latent multi-view subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4279–4287, 2017.