GCFAgg: Global and Cross-view Feature Aggregation for Multi-view Clustering

Weiqing Yan[1,5], Yuanyang Zhang[1], Chenlei Lv[2], Chang Tang[3]*, Guanghui Yue[4], Liang Liao[5], Weisi Lin[5]

[1]School of Computer and Control Engineering, Yantai University, Yantai 264005, China

[2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

[3] School of Computer, China University of Geosciences, Wuhan 430074, China

[4] School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China

[5] School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore.

Abstract

Multi-view clustering can partition data samples into their categories by learning a consensus representation in unsupervised way and has received more and more attention in recent years. However, most existing deep clustering methods learn consensus representation or view-specific representations from multiple views via view-wise aggregation way, where they ignore structure relationship of all samples. In this paper, we propose a novel multi-view clustering network to address these problems, called Global and Cross-view Feature Aggregation for Multi-View Clustering (GCFAggMVC). Specifically, the consensus data presentation from multiple views is obtained via cross-sample and cross-view feature aggregation, which fully explores the complementary of similar samples. Moreover, we align the consensus representation and the view-specific representation by the structure-guided contrastive learning module, which makes the view-specific representations from different samples with high structure relationship similar. The proposed module is a flexible multi-view data representation module, which can be also embedded to the incomplete multi-view data clustering task via plugging our module into other frameworks. Extensive experiments show that the proposed method achieves excellent performance in both complete multi-view data clustering tasks and incomplete multi-view data clustering tasks.

1. Introduction

With the rapid development of informatization, data is often collected by various social media or views. For instance, a 3D object can be described from different angles; a news event is reported from different sources; and an image can be characterized by different types of feature sets, e.g., SIFT, LBP, and HoG. Such an instance object, which is described from multiple views, is referred to as multi-view data. Multi-view clustering (MVC) [6], i.e., unsupervisedly fusing the multi-view data to aid differentiate crucial grouping, is a fundamental task in the fields of data mining, pattern recognition, etc, but it remains a challenging problem.

Traditional multi-view clustering methods [7] include matrix decomposition methods, graph-based multi-view methods, and subspace-based multi-view methods. The goal of these methods is to obtain a high-quality consensus graph or subspace self-representation matrix by various regularization constraints in order to improve the performance of clustering. However, most of them directly operate on the original multiview features or specified kernel features, which usually include noises and redundancy information during the collection or kernel space selection processes, moreover harmful to the clustering tasks.

Deep neural networks have demonstrated excellent performance in data feature representation for many vision tasks. Deep clustering methods also draw more attention to researchers [1, 9, 40, 50–52]. These methods efficiently learn the feature presentation of each view using a view-specific encoder network, and fuse these learnt representations from all views to obtain a consensus representation that can be divided into different categories by a clustering module. To reduce the influence of view-private information on clustering, these methods designed different alignment models. For example, some methods align the representation distributions or label distributions from different views by KL divergence [14]. They might be hard to distinguish between clusters, since a category from one view might be aligned with a different category in another view. Some methods align the representation from different views by contrastive learning. Despite these models have achieved significant improvement in MVC task, the following issues still exist: 1) Almost all existing deep MVC methods (such as [38,40,49]) are based on view-wise fusion models, such as weighted-sum fusion of all views or concatenating fusion of all views, which makes it difficult to

---

*Corresponding author(tangchang@cug.edu.cn)

obtain discriminative consensus representations from multiple views, since a view or several views of a sample might contain too much noise or be missing in the collection process. 2) These alignment methods (such as [38, 51]) based on contrastive learning usually distinguish the positive pair and negative pair from the sample-level. That is, they make inter-view presentations from the same sample as positive pair, and makes view representations from different samples as negative pair (including view representations from different samples in the same cluster), whereas it might be conflict with the clustering objective, where these representations of samples from the same cluster should be similar to each other.

In the paper, a novel multi-view representation learning framework for clustering is proposed to alleviate the above problems. Motivated by the insight that the representations of samples from the same category are typically similar, we can learn consensus data representation from multiview data by other samples with a high structure relationship, moreover, in contrastive learning, we should increase the similarity of view representations from the same cluster, not only from the same sample, which is beneficial clustering tasks. To accomplish this, we first learn the view-specific representations to reconstruct the original data by leveraging the autoencoder model. Then, we design a global and cross-view feature aggregation module, which is capable of learning a global similarity relationship among samples, and obtaining a consensus representation based on the global similarity relationship of all samples. Furthermore, we leverage the learnt global structure relationship and consensus representation to establish the consistency with view-specific representations by contrastive learning, which minimizes the similarity between the representations with low structure relationship. Compared with previous work, our contributions are listed as follows:

- We propose a novel Global and Cross-view Feature Aggregation network framework for Multi-View Clustering (GCFAggMVC), which is able to fully explore the complementary of similar samples and addresses the problem of negative pairs from the different samples in the same cluster having low similarity score.

- Different from previous methods, we design a global and cross-view feature aggregation module, which integrates the transformer structure to learn the global structure relationship from different feature spaces, and then obtains the consensus representation based on the learnt global relationship, which fully exploits the complementary information of similar samples, thereby reduce the impact of noise and redundancy or sample missing among different views. Moreover, we align the consensus representation and the view-specific representation by our global structure-guided

contrastive learning module, which makes the representations of similar samples with highly structure relationship similarity.

- The proposed module is flexible multi-view data representation module, which can be also applied to incomplete multi-view data clustering tasks by plugging our module into the framework of other methods as the consensus representation of a sample with missing view data can be enhanced by these samples with high structure relationships. Experiments show that the proposed method achieves not only excellent performance in complete multi-view clustering tasks, but also works well in incomplete multi-view clustering tasks.

2. RETATED WORK

In the existing representation learning-based MVC family [7], there are mainly two kinds of methods, i.e., shallow representation learning-based method and deep representation learning-based method. The following section, a brief review will be introduced.

2.1. Shallow representation learning-based method

Shallow representation learning-based multi-view clustering methods [7] are divided into the two categories, multi-view graph clustering and multi-view subspace clustering.

For multiview graph clustering [11, 16, 19, 28, 32, 33], they generally do the following steps: construct view-specific graphs, and then obtain a fusion graph by different regularization terms from multiple view-specific graphs, last, produce the clustering results by spectral clustering or graph-cut methods or other algorithms.

For subspace-based multiview clustering [5, 13, 20, 30, 37, 45], they learn a consensus subspace self-representation matrix from each view, and partition the data points into different subspaces by applying different regularization terms on the consensus self-representation matrix. To reducing the algorithm complexity, some methods [20, 37] computed similarity graph or self-expression depending on the bipartite graph between samples and anchors instead of all samples-based graph methods. These methods directly learn graph or self representation in the original feature space in which redundant and noisy features of samples are inevitably mixed. Some methods [10, 25, 29] map raw feature space to high-dimensional feature space by kernelization method, however, these methods assume that the raw data can be represented by one or fixed few kernel matrices, which might be insufficient to describe complex data.

2.2. Deep representation learning-based method

Motivated by the promising progress of deep learning in powerful feature transformation ability, many recent works
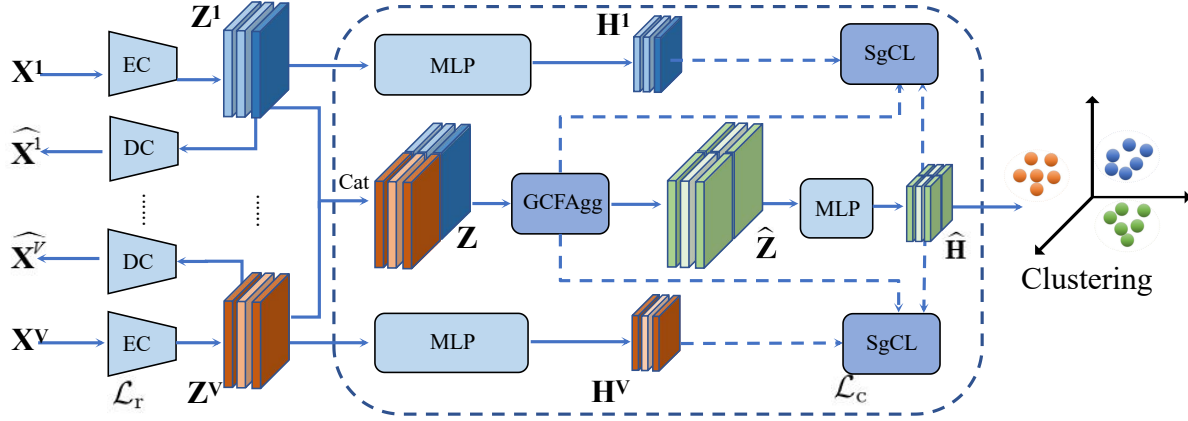
Figure 1. The overall framework. Our module includes global and cross-view feature aggregation module (GCFAgg) and structure-guided multiview contrastive learning module (SgCL). The former learns a consensus representation via considering global structure relationship among samples, which fully explores the complementary of similar samples. The latter integrates the learnt global structure relationship and consensus representation to contrastive learning, which makes data representations in the same cluster similar and addresses the aforementioned second issue in the introduction. Note that, EC: Encoder; DC: Decoder; Cat: Concatenation; MLP: Multi-Layer Perception.

have been focused on the deep representation learning-based multi-view clustering. Specifically, these methods use the deep neural network to model the non-linear parametric mapping functions, from which the embedded feature representation canbe learnt the manifold structure by the non-linearity properties.

Some deep multi-view clustering methods [24,26,44,52] use adversarial training to learn the latent feature representations and align distributions of hidden representations from different views. Zhou et al. [52] leveraged the attention mechanism to obtain a weight value for each view, and obtained a consensus representation by weight summation of all view-wise presentations. Wang et al. [44] obtained the consensus representation by a weight summation and $l_{1,2}$-norm constraint. Contrastive learning (CL) is able to align representations from different views at the sample level, further facilitating the alignment of label distributions as well. These CL methods [8, 21, 27, 33, 39, 40, 51] have exceeded the previous distribution alignment methods to multi-view clustering, yielding SOTA clustering performance on several multi-view datasets.

3. PROPOSED METHOD

The proposed framework is shown in Fig. 1. In this paper, a multi-view data, which includes n samples with V views, is denoted as $\{\mathbf{X}^v = \{\mathbf{x}_1^v; ...; \mathbf{x}_N^v\} \in \mathbb{R}^{N \times D_v}\}_{v=1}^V$, where D_v is the feature dimension in the v-th view.

3.1. Multi-view Data Reconstruction

Original multi-view data usually contains redundancy and random noise, we need to first learn representative feature representations from the original data features. In particular, autoencoder [15, 36] is a widely used unsupervised

model that projects original data features into a designed feature space. Specifically, for the v-th view, let $f_{\theta^v}^v(.)$ represent the encoder nonlinear function. In the encoder, the network learns the low-dimensional features as follows:

$$\mathbf{z}_i^v = f_{\theta^v}^v(\mathbf{x}_i^v), \tag{1}$$

where $\mathbf{z}_i^v \in \mathbb{R}^{d_v}$ is the embedded data representation in d_v-dimensional feature space of the i-th sample from the v-th view $\mathbf{x}_i^v$.

The decoder reconstructs the sample by the data representation $\mathbf{z}_i^v$. Let $g_{\eta^v}^v(.)$ represent the decoder function, in decoder part, the reconstructed sample $\hat{\mathbf{x}}_i^v$ is obtained by decoding $\mathbf{z}_i^v$:

$$\hat{\mathbf{x}}_i^v = g_{\eta^v}^v(\mathbf{z}_i^v) = g_{\eta^v}^v(f_{\theta^v}^v(\mathbf{x}_i^v)) \tag{2}$$

Let $\mathcal{L}_r$ be the reconstruction loss from input to output $\hat{\mathbf{X}}^v = \{\hat{\mathbf{x}}_1^v; ...; \hat{\mathbf{x}}_n^v\} \in \mathbb{R}^{n \times D_v}$, n denotes the number of samples in a batch. The reconstruction loss is formulated as:

$$\mathcal{L}_r = \sum_{v=1}^V \mathcal{L}_r^v = \sum_{v=1}^V \left\| \mathbf{X}^v - \hat{\mathbf{X}}^v \right\|_2^2$$
$$= \sum_{v=1}^V \sum_{i=1}^n \left\| \mathbf{x}_i^v - g_{\eta^v}^v(\mathbf{z}_i^v) \right\|_2^2 \tag{3}$$

3.2. Global and Cross-view Feature Aggregation

In Eq. (3), to reduce the loss of reconstruction, the learned data representation $\mathbf{z}_i^v$ usually contains much view-private information, which may be meaningless or even misleading. They might result in poor clustering quality if they are fused by previous view-wise fusion methods.

Since the representations of different samples in the same cluster are similar, the consensus representation of a
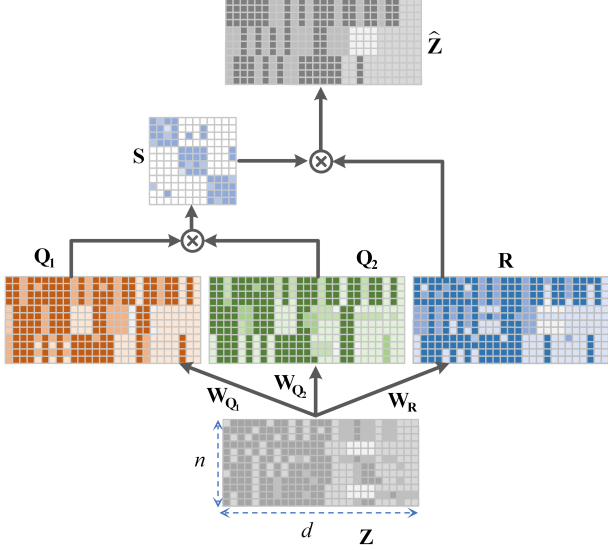
Figure 2. Global and Cross-view Feature Aggregation module . $\mathbf{W}_{Q_1}, \mathbf{W}_{Q_2}, \mathbf{W}_R$ is used to achieve feature space transformation of cross-view, and the $\mathbf{S}$ is used to obtain global structure relationship among samples.

sample from different views should be enhanced by those samples with strong structure relationships, not just the weight-sum of representation of different views from the same sample. Therefore, to obtain consensus multi-view data representation, we learn the global structure relationship among samples and use it to obtain the consensus data representation. The module is shown in Figure 2. Specifically, we firstly concatenate all view-specific representations $\mathbf{Z}^v$ extracted by the encoder together to get $\mathbf{Z}$ is formulated as:

$$\mathbf{Z} = \left[\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^V\right] \tag{4}$$

where $\mathbf{Z}^v = [\mathbf{z}_1^v; \mathbf{z}_2^v; \ldots; \mathbf{z}_n^v] \in \mathbb{R}^{n \times d_v}$, $\mathbf{z}_i^v \in \mathbb{R}^{1 \times d_v}$, $\mathbf{Z} \in \mathbb{R}^{n \times d}$, $d = d_v * V$. Here, we set each sample as a token, set linear encoding and position encoding.

Inspired by the idea of the transformer attention mechanism [42], we map $\mathbf{Z}$ into different feature spaces by the $\mathbf{W}_R$ to achieve the cross-view fusion of all views, i.e.,

$$\begin{bmatrix} \mathbf{R}_{1:} \\ \mathbf{R}_{2:} \\ \vdots \\ \mathbf{R}_{n:} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_1^1 & \mathbf{z}_1^2 & \cdots & \mathbf{z}_1^V \\ \mathbf{z}_2^1 & \mathbf{z}_2^2 & \cdots & \mathbf{z}_2^V \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}_n^1 & \mathbf{z}_n^2 & \cdots & \mathbf{z}_n^V \end{bmatrix} \begin{bmatrix} \mathbf{W}_{R1:} \\ \mathbf{W}_{R2:} \\ \vdots \\ \mathbf{W}_{RV:} \end{bmatrix} \tag{5}$$

that is $\mathbf{R}_{j:} = \sum_{k=1}^{V} \mathbf{z}_j^k \mathbf{W}_{Rk:}$. Similarity, the $\mathbf{Q}_1$ and $\mathbf{Q}_2$ is obtained by $\mathbf{W}_{Q_1}, \mathbf{W}_{Q_2}$, i.e.,

$$\mathbf{Q}_1 = \mathbf{Z}\mathbf{W}_{Q_1}; \mathbf{Q}_2 = \mathbf{Z}\mathbf{W}_{Q_2}; \tag{6}$$

where $\mathbf{Q}_1 \in \mathbb{R}^{n \times d}$, $\mathbf{Q}_2 \in \mathbb{R}^{n \times d}$. Here, we use the matrix $\mathbf{W_Z} = \{\mathbf{W}_{Q_1}, \mathbf{W}_{Q_2}, \mathbf{W}_R\}$ to denote the parameters in this module.

The structure relationship among samples is denoted as:

$$\mathbf{S} = \text{softmax}\left(\frac{\mathbf{Q}_1\mathbf{Q}_2^T}{\sqrt{d}}\right). \tag{7}$$

The learned representation matrix $\mathbf{R}$ is enhanced by the global structure relationship matrix $\mathbf{S}$. That is, the data representation of the sample can be enhanced by other samples with high correlation, which makes these data representations of samples in the same cluster be similar.

$$\widehat{\mathbf{z}_i} = \sum_{j=1}^{n} \mathbf{S}_{ij}\mathbf{R}_{j:}; \hat{\mathbf{Z}} = [\widehat{\mathbf{z}_1}; \widehat{\mathbf{z}_2}; \ldots; \widehat{\mathbf{z}_n}] \tag{8}$$

where $\mathbf{R}_{j:} \in \mathbb{R}^{1 \times d}$ is the j-th row elements of $\mathbf{R}$, denotes the j-th sample representation, $\mathbf{S}_{ij}$ denotes the relationship between the i-th sample and the j-th sample, $\hat{\mathbf{Z}} \in \mathbb{R}^{n \times d}$. Since $\hat{\mathbf{Z}}$ is learnt from the concatenation of all views $\mathbf{Z}$, it usually contains redundancy information. Next, the output is passed through the fully connected nonlinear and linear layer to eliminate the redundancy information. The expression is described as the following equation:

$$\widehat{\mathbf{H}} = \mathbf{W_3}\left(\max(0, (\mathbf{Z} + \hat{\mathbf{Z}})\mathbf{W_1} + \mathbf{b_1})\mathbf{W_2} + \mathbf{b_2}\right) + \mathbf{b_3} \tag{9}$$

where we use $\mathbf{W}_{\widehat{\mathbf{H}}}$ to denote these parameters in the layers.

3.3. Structure-guided Contrastive Learning

The learnt consensus representation $\hat{\mathbf{H}}$ is enhanced by global structure relationship of all samples in a batc, these data consensus representations from different views of samples in the same cluster are similar. Hence, the consensus representation $\hat{\mathbf{H}}$ and view-specific representation $\mathbf{H}^v$ from the same cluster should be mapped close together. Inspired by contrastive learning methods [8], we set the consensus representation and view-specific representation from different views from the same sample as positive pairs. However, if other pairs are directly set as negative pairs, it might lead to the inconsistency of these represents from different samples in the same cluster, which is conflict with clustering objective. Hence, we design a structure-guided multiview contrastive learning module. Specifically, we introduce cosine distance to compute the similarity between consensus presentation $\hat{\mathbf{H}}$ and view-specific presentation $\mathbf{H}^v$:

$$C\left(\hat{\mathbf{H}}_{i:}, \mathbf{H}_i^v\right) = \frac{\hat{\mathbf{H}}_{i:}^T\mathbf{H}_{i:}^v}{\left\|\hat{\mathbf{H}}_{i:}\right\|\left\|\mathbf{H}_{i:}^v\right\|} \tag{10}$$

The loss function of structure-guided multi-view contrastive learning can be defined as:

$$\mathcal{L}_c = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{v=1}^{V}\log\frac{e^{C(\hat{\mathbf{H}}_{i:}, \mathbf{H}_{i:}^v)/\tau}}{\sum_{j=1}^{N}e^{(1-\mathbf{S}_{ij})C(\hat{\mathbf{H}}_{i:}, \mathbf{H}_{j:}^v)/\tau} - e^{1/\tau}} \tag{11}$$

where τ denotes the temperature parameter, $\mathbf{S}_{ij}$ is from Eq. (7). In this equation, when the smaller this $\mathbf{S}_{ij}$ value is, the bigger the $\mathrm{C}\left(\widehat{\mathbf{H}}_{i:}, \mathbf{H}_{j:}^{v}\right)$. In other words, when the structure relationship $\mathbf{S}_{ij}$ between the i-th and j-th sample is low (not from the same cluster), their corresponding representations are inconsistent; otherwise, their corresponding representations are consistent, which solves the aforementioned second issue in the introduction.

In the proposed framework, the loss in our network consists of two parts:

$$
\begin{aligned}
\mathcal{L} &= \mathcal{L}_{\mathrm{r}} + \lambda \mathcal{L}_{\mathrm{c}} \\
&= \mathcal{L}_{\mathrm{r}}\left(\{\mathbf{X}^{v}, \widehat{\mathbf{X}}^{v}\}_{v=1}^{V}; \{\eta^{v}, \theta^{v}\}_{v=1}^{V}\right) + \\
&\quad \lambda \mathcal{L}_{\mathrm{c}}\left(\{\mathbf{Z}, \widehat{\mathbf{H}}, \mathbf{H}^{v}\}_{v=1}^{V}; \{\mathbf{W}_{\mathbf{Z}}, \mathbf{W}_{\widehat{\mathbf{H}}}, \mathbf{W}_{\mathbf{H}^{v}}, \theta^{v}\}_{v=1}^{V}\right)
\end{aligned}
\tag{12}
$$

where $\mathcal{L}_{\mathrm{c}}$ is the reconstruction loss from input $\mathbf{X}^{v}$ to the output $\widehat{\mathbf{X}}^{v}$ by the embedding representation matrix $\mathbf{Z}^{v}$, η^{v}, θ^{v} denote the parameters of encoder and decoder, this loss is used to avoid the model collapse. $\mathcal{L}_{\mathrm{r}}$ is the consistent data presentation loss by contrastive learning between structure-enhanced consensus data representation $\widehat{\mathbf{H}}$ and view-specific data representation $\mathbf{H}^{v}$.

Different from previous contrastive learning methods, which align inter-view representations, our method has the following advantages: 1) we add the structure relationship $\mathbf{S}$ (Eq. (7)) obtained by feature fusion to negative pairs, which ensures that we only minimize the similarity between the view-specific representation and the consensus representation from different samples with low structure relationship. 2) for the positive pairs, the proposed method aligns the learnt consensus representation and view-specific representation, which makes the representations of positive pair with high structure relationship be more similar, since the consensus presentation is enhanced with by other samples with high correlation. The proposed method breaks the limitations of previous CL at sample level.

3.4. Clustering module

For the clustering module, we take the k-means [2,31] to obtain the clustering results for all samples. Specifically, the learnt consensus representation $\widehat{\mathbf{H}}$ is factorized as follows:

$$
\begin{aligned}
&\min_{\mathbf{U},\mathbf{V}} \quad \left\|\widehat{\mathbf{H}} - \mathbf{U}\mathbf{V}\right\|^{2} \\
&s.t. \mathbf{U}\mathbf{1} = \mathbf{1}, \mathbf{U} \geq \mathbf{0}
\end{aligned}
\tag{13}
$$

where $\mathbf{U} \in \mathbb{R}^{n \times k}$ is cluster indicator matrix; $\mathbf{V} \in \mathbb{R}^{k \times d}$ is the center matrix of clustering.

3.5. Optimization

The model consists of multiple encoder-decoder modules and multiple MLP layers. The model is optimized by a mini-batch gradient descent algorithm. Firstly, the autoencoders are initialized by Eq. (3), and structure-guided contrastive learning are conducted to obtain the consensus representation by Eq. (12). At last, the cluster labels are obtained by Eq. (13).

4. Experiments

4.1. Experimental Settings

We evaluate our models on 13 public multi-view datasets with different scales (see Table 1). For the evaluation metrics, three metrics, including accuracy (ACC), normalized mutual information (NMI), and Purity (PUR).

Table 1. Description of the multiview datasets.

Datasets	Samples	Views	Clusters
Prokaryotic [3]	551	3	4
Synthetic3d [23]	600	3	3
MNIST-USPS [35]	5000	2	10
CCV [18]	6773	3	20
Hdigit [7]	10000	2	10
Cifar10 [1]	50000	3	10
Cifar100 [2]	50000	3	100
YouTubeFace[3]	101499	5	31
Caltech-5V [12]	1400	5	7
NGs [4]	500	3	5
Cora [47]	2708	2	7
BDGP [4]	2500	2	5
Fashion [48]	10000	3	10

Compared methods: To evaluate the effectiveness of the proposed method, we compare the proposed method with 13 SOTA clustering methods, including 9 methods for complete multi-view datasets and 4 methods for incomplete multi-view datasets. The former includes 4 traditional methods (PLCMF [43], LMVSC [20], SMVSC [37], and FastMICE [17]) and 5 deep methods (DEMVC [50], CO-NAN [21], SiMVC [40], CoMVC [40], MFLVC [51]). The latter includes CDIMC [46], COMPLETER [27], DIMVC [49], DSIMVC [38].

Implementation Details: The experiments were conducted on Linux with i9-10900K CPU, 62.5GB RAM and 3090Ti GPU. We pre-train the models for 200 epochs for the reconstruction loss and then fine-tune the model for 100 epochs on mini-batches of size 256 using the Adam optimizer [22] in the PyTorch [34] framework. In the proposed method, we reshape all datasets into vectors and implement autoencoders using a fully connected network.

---

[1] http://www.cs.toronto.edu/kriz/cifar.html
[2] http://www.cs.toronto.edu/kriz/cifar.html
[3] https://www.cs.tau.ac.il/ wolf/ytfaces/
[4] https://lig-membres.imag.fr/grimal/data.html

Table 2. Clustering result comparison for different datasets.

Datasets	CCV			MNIST-USPS			Prokaryotic			Synthetic3d		
Metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
PLCMF [43]	0.2294	0.1852	0.2557	0.6228	0.6594	0.6670	0.4446	0.0200	0.5681	0.8867	0.6555	0.8867
LMVSC [20]	0.2014	0.1657	0.2396	0.5626	0.5039	0.6060	0.5753	0.1337	0.6294	0.9567	0.8307	0.9567
SMVSC [37]	0.2182	0.1684	0.2439	0.7542	0.6883	0.7542	0.5590	0.1820	0.5717	0.9683	0.8665	0.9683
FastMICE [17]	0.1997	0.1518	0.2341	0.9570	0.9332	0.9573	0.5629	0.2685	0.6500	0.9613	0.8490	0.9613
DEMVC [50]	0.1942	0.2113	0.2169	0.8858	0.9100	0.8880	0.5245	0.3079	0.6969	0.8100	0.6136	0.8100
CONAN [21]	0.1422	0.1016	0.1674	0.5722	0.5708	0.6178	0.4809	0.1589	0.5045	0.9650	0.8540	0.9650
SiMVC [40]	0.1513	0.1252	0.2161	0.9810	0.9620	0.9810	0.5009	0.1945	0.6098	0.9366	0.7747	0.9366
CoMVC [40]	0.2962	0.2865	0.2976	0.9870	0.9760	0.9890	0.4138	0.1883	0.6697	0.9530	0.8184	0.9520
MFLVC [51]	0.3123	0.3162	0.3391	0.9954	0.9869	0.9898	0.4301	0.2216	0.5989	0.9650	0.8537	0.9650
Ours	**0.3543**	**0.3292**	**0.3812**	**0.9956**	**0.9871**	**0.9956**	**0.6225**	**0.3778**	**0.7314**	**0.9700**	**0.8713**	**0.9700**

Table 3. Clustering result comparison for different datasets.

Datasets	Hdigit			YouTubeFace			Cifar10			Cifar100		
Metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
PLCMF [43]	0.9047	0.7965	0.9047	0.1473	0.1237	0.2875	0.8144	0.8265	0.8497	0.8260	0.9593	0.8698
LMVSC [20]	0.9709	0.9293	0.9709	0.1479	0.1327	0.2816	0.9896	0.9721	0.9896	0.8482	0.9583	0.9582
SMVSC [37]	0.8634	0.7683	0.8634	0.2587	0.2292	0.3321	0.9899	0.9730	0.9899	0.7429	0.9091	0.7529
FastMICE [17]	0.9332	0.9258	0.9417	0.1825	0.1633	0.3028	0.9694	0.9622	0.9704	0.8257	0.9464	0.8298
DEMVC [50]	0.3738	0.3255	0.4816	0.2487	0.0932	0.2662	0.4354	0.3664	0.4498	0.5048	0.8343	0.5177
CONAN [21]	0.9562	0.9193	0.9562	0.1179	0.1178	0.1499	0.9255	0.8641	0.9255	0.6711	0.9441	0.9983
SiMVC [40]	0.7854	0.6705	0.7854	0.0765	0.0481	0.2662	0.8359	0.7324	0.8359	0.5795	0.9225	0.5869
CoMVC [40]	0.9032	0.8713	0.9032	0.1010	0.0851	0.2674	0.9275	0.8925	0.9275	0.6569	0.9345	0.6570
MFLVC [51]	0.9442	0.8750	0.9440	0.2770	0.2952	0.3297	0.9918	0.9774	0.9918	0.8268	0.9560	0.8268
Ours	**0.9744**	**0.9305**	**0.9744**	**0.3262**	**0.3289**	**0.4007**	**0.9923**	**0.9781**	**0.9923**	**0.9597**	**0.9935**	**0.9605**

For the comparison with these methods on incomplete multi-view datasets, the incomplete samples is ready according to the method [38] on Synthetic3D, NGs, Cora, BDGP, and Fashion by randomly removing views under the condition that at least one view remained in the sample. The ratio of incomplete sample sizes to overall sample sizes is set from 0.1 to 0.7 with 0.2 as the interval. Please refer to the supplementary material for the experiment details for the proposed method in incomplete datasets.

4.2. Experimental comparative results

The comparative results with 9 methods by three evaluation metrics (ACC, NMI, PUR) on 8 benchmark datasets with different scales are presented as Table 2 and Table 3. From the tables, we can observe that the proposed GCFAggMVC obtains better results than those of other methods. Specifically, we obtain the following observations:

(1) Our method obtains better results than four traditional multiview clustering methods (PLCMF, LMVSC, SMVSC, and FastMICE). These methods learn the data self-representation or graph structure relationship on the raw data features. These raw data features contain noises and redundancy information, which is harmful to establish the essential graph structure. From this table, it can be seen that our method achieves better performance.

(2) We compare five deep multiview clustering methods (DEMVC, CONAN, SiMVC, CoMVC, and MFLVC) with the proposed method. In DEMVC, they consider to align label distributions of different views to a target distribution, which has a negative impact since a given cluster distribution from one view might be aligned with a different cluster distribution from another view [40]. In CONAN, SiMVC, and CoMVC, they obtain a consensus feature representation by view-wise fusion method. However, the meaningless view-private information might be dominant in the feature fusion, and thus harmful to the clustering performance. Compared with these methods, in the proposed method, the consensus representation of each sample from different views can be enhanced by those samples with high structure relationship, moreover, maximizes the consistency of the view-specific representations and consensus data representation to obtain efficient clustering performance. From this table, it can be seen that GCFAggMVC obtains better results than those of other methods.

To further verify the effectiveness of the proposed method, we test the performance of different numbers of views on the Caltech dataset. Table 4 shows the comparative results with different deep clustering methods. From this table, we can see that the results of the proposed GCFAggMVC outperform those of the other methods.

Table 4. Comparisons with deep clustering methods on Caltech dataset with increased views. "XV" denotes the number of views.

Datasets	Caltech-2V			Caltech-3V			Caltech-4V			Caltech-5V		
Metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
DEMVC [50]	0.4986	0.3845	0.5207	0.5336	0.4136	0.5336	0.4929	0.4504	0.5186	0.4600	0.3666	0.4950
CONAN [21]	0.5750	0.4516	0.5757	0.5914	0.4981	0.5914	0.5571	0.5061	0.5735	0.7207	0.6418	0.7221
SiMVC [40]	0.5083	0.4715	0.5573	0.5692	0.4953	0.5912	0.6193	0.5362	0.6303	0.7193	0.6771	0.7292
CoMVC [40]	0.4663	0.4262	0.5272	0.5413	0.5043	0.5842	0.5683	0.5692	0.6463	0.7003	0.6871	0.7462
MFLVC [51]	0.6060	**0.5280**	0.6160	0.6312	**0.5663**	0.6392	0.7332	0.6523	0.7342	0.8042	0.7032	0.8043
Ours	**0.6643**	0.5008	**0.6643**	**0.6400**	0.5345	**0.6529**	**0.7343**	**0.6610**	**0.7343**	**0.8336**	**0.7331**	**0.8336**

Table 5. Clustering results on incomplete datasets. "-" denotes unknown results as COMPLETER mainly focuses on two-view clustering.

Dataset	Missing rates	0.1			0.3			0.5			0.7		
	Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
BDGP	CDIMC [46]	0.8047	0.7008	0.8037	0.7467	0.6764	0.7527	0.6771	0.5451	0.6771	0.5611	0.3970	0.5776
	COMPLETER [27]	0.4091	0.4180	0.4154	0.3963	0.3319	0.3115	0.3262	0.2747	0.4390	0.4359	0.4510	0.4090
	DIMVC [49]	0.9640	0.8920	0.9120	0.9540	0.8660	0.8890	0.9470	0.8450	0.8730	0.9290	0.8020	0.8310
	DSIMVC [38]	0.9827	0.9443	0.9827	0.9693	0.9034	0.9693	0.9529	0.8611	0.9529	0.9214	0.7937	0.9214
	DSIMVC++ (Our)	**0.9836**	**0.9455**	**0.9836**	**0.9698**	**0.9050**	**0.9698**	**0.9557**	**0.8685**	**0.9557**	**0.9332**	**0.8142**	**0.9332**
Synthetic3d	CDIMC [46]	0.5965	0.3564	0.6000	0.5124	0.2239	0.5340	0.5330	0.2373	0.5663	0.4136	0.1387	0.4394
	COMPLETER [27]	-	-	-	-	-	-	-	-	-	-	-	-
	DIMVC [49]	**0.8183**	0.6701	0.8380	**0.8233**	0.5860	0.8241	**0.7968**	0.5355	0.7971	0.6689	0.3974	0.6774
	DSIMVC [38]	0.7613	0.6744	0.8943	0.7378	0.6365	0.8773	0.7247	0.6090	0.8643	0.7043	0.5499	0.8242
	DSIMVC++ (Our)	0.7785	**0.6933**	**0.9005**	0.7530	**0.6693**	**0.9042**	0.7612	**0.6463**	**0.8952**	**0.7197**	**0.5900**	**0.8638**
Cora	CDIMC [46]	0.2460	0.0111	0.3066	0.2222	0.0066	0.3024	0.2400	0.0052	0.3022	0.2518	0.0054	0.3025
	COMPLETER [27]	0.2441	**0.4300**	0.3172	0.2542	**0.4130**	0.3242	0.2464	**0.4070**	0.3199	0.2540	0.1850	0.3055
	DIMVC [49]	0.4384	0.2231	0.5079	0.3704	0.1470	0.4082	0.3561	0.1432	0.4275	0.2789	0.0718	0.3397
	DSIMVC [38]	0.4402	0.3316	0.5445	0.4106	0.2924	0.5099	0.3764	0.2360	0.4742	0.3243	0.1628	0.4228
	DSIMVC++ (Our)	**0.4699**	0.3271	**0.5588**	**0.4484**	0.3035	**0.5544**	**0.4338**	0.2720	**0.5290**	**0.3554**	**0.1935**	**0.4620**
NGs	CDIMC [46]	0.3072	0.0794	0.3216	0.2736	0.0478	0.2832	0.2532	0.0346	0.2620	0.2464	0.0270	0.2504
	COMPLETER [27]	-	-	-	-	-	-	-	-	-	-	-	-
	DIMVC [49]	0.3543	0.1493	0.3562	0.2120	0.0363	0.2138	0.2213	0.0588	0.2280	0.2598	0.0546	0.2645
	DSIMVC [38]	0.5564	0.4599	0.6230	0.5178	0.3864	0.5854	**0.4672**	**0.2980**	0.5244	0.4090	0.2095	0.4746
	DSIMVC++ (Our)	**0.6358**	**0.5186**	**0.7090**	**0.6136**	**0.4428**	**0.6734**	0.4598	0.2801	**0.5310**	**0.4410**	**0.2298**	**0.5054**
Fashion	CDIMC [46]	0.6500	0.6642	0.6696	0.5064	0.5121	0.5241	0.4484	0.4483	0.4553	0.3693	0.3668	0.3818
	COMPLETER [27]	-	-	-	-	-	-	-	-	-	-	-	-
	DIMVC [49]	0.7811	0.8578	0.8286	0.7132	0.7676	0.7614	0.7044	0.7447	0.7508	0.6128	0.6806	0.6693
	DSIMVC [38]	0.8798	0.8623	0.8800	0.8680	0.8379	0.8687	0.8333	0.8025	0.8337	0.7825	0.7626	0.7825
	DSIMVC++ (Our)	**0.9360**	**0.8953**	**0.9360**	**0.9160**	**0.8657**	**0.9160**	**0.8969**	**0.8366**	**0.8969**	**0.8637**	**0.8015**	**0.8644**

Since our consensus representation is enhanced by those samples with high structure relationships, the proposed GCFAgg module and SgCL loss canbe plugged into incomplete multi-view clustering task. To verify the effectiveness, we integrate the proposed GCFAgg module and SgCL loss to the state-of-the-art DSIMVC (which is an incomplete multi-view clustering method) framework (named DSIMVC++, please refer to the appendix for the details) and compare it with the other 4 methods in Table 5. In the DSIMVC++ method, the consensus feature presentation for clustering can be enhanced by those samples with high structure relationship, hence, the propose method can better cope with incomplete multiview clustering tasks. From the table, it can be observed that our GCFAggMVC achieves better performance than other methods. Especially, when the missing view rate is 0.7, it can be observed that the proposed GCFAggMVC can obtain the best results than other all comparison methods.

4.3. Model analysis

Convergence analysis. To verify the convergence, we plot the objective values and evaluation metric values through iterations in Figure 3. It can be observed that the objective value monotonically decreases until convergence. The value of ACC, NMI, and PUR first increases gradually with iteration and then fluctuate in a narrow range. These results all confirm the convergence of GCFAggMVC.

Parameter sensitivity analysis. We experimentally evaluate the effect of hyperparameters on the clustering performance of GCFAggMVC, which includes the trade-off coefficient λ (i.e., $\mathcal{L}_r + \lambda \mathcal{L}_c$) and the temperature parameter τ. Figure 3 shows the NMI of GCFAggMVC when λ is varied from 10^{-3} to 10^3 and τ from 0.2 to 0.8. From this figure, the clustering results of the proposed GCFAggMVC are insensitive to both λ and τ in the range 0.1 to 1, and the range 0.3 to 0.5, respectively. Empirically, we set λ and τ to 1.0 and 0.5.
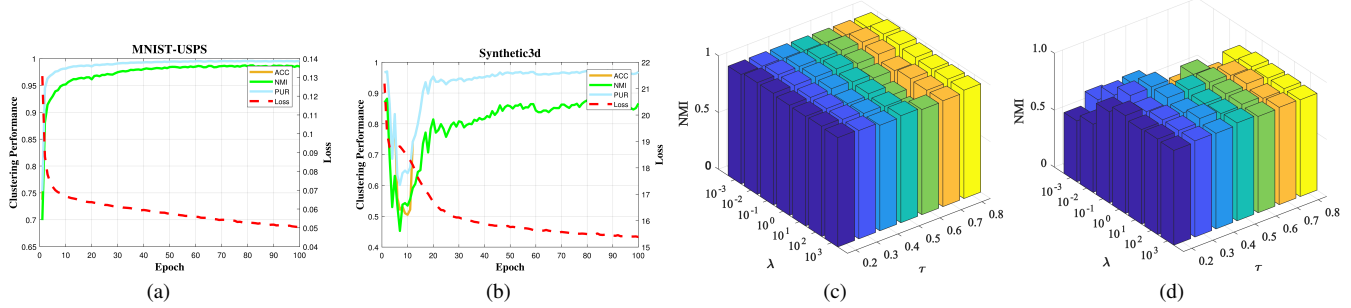
Figure 3. The convergence analysis and parameter analysis on MNIST-USPS, Synthetic3d, respectively.

Table 6. Ablation study.

Datasets	Method	ACC	NMI	PUR
Prokaryotic	No-GCFAgg	0.4403	0.1906	0.5740
	No-SgCL	0.4804	0.2226	0.6534
	GCFAggMVC	**0.6225**	**0.3778**	**0.7314**
CCV	No-GCFAgg	0.2850	0.2740	0.3150
	No-SgCL	0.2020	0.1900	0.2560
	GCFAggMVC	**0.3543**	**0.3292**	**0.3812**
MNIST-USPS	No-GCFAgg	0.9753	0.9500	0.9753
	No-SgCL	0.6949	0.6656	0.7410
	GCFAggMVC	**0.9956**	**0.9871**	**0.9956**

(a) **Z** features   (b) **H** features   (c) **Ĥ** features
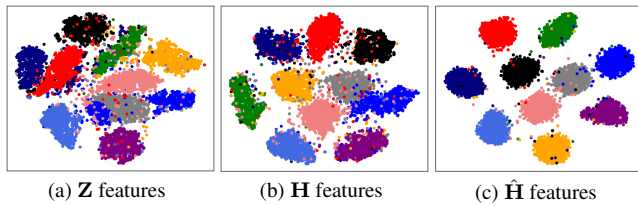
Figure 4. The visualization results of different feature representations on different layers after convergence. Note that, **H** feature denotes the concatenation of all learnt $\mathbf{H}^v$.

Ablation Study. We conducted an ablation study to evaluate each component of the proposed model. We perform two models and compare them with the proposed method.

Validity of the proposed GCFAgg module: we set the **Z** (concatenation of all view-specific representations) as the consensus representation. This method is denoted as "No-GCFAgg". As shown in Table 6, it can be observed that the results of No-GCFAgg are lower than the results of the proposed method by 18.22, 6.93, and 2.03 percent in ACC term. The concatenated representation **Z** includes much view-private information, which is not conductive to clustering. The GCFAgg fully explores the complementary of similar samples, thereby reducing the impact of noise and redundancy among all views. The results show that our GCFAgg module can enhance the consistency of data representation from the same cluster.

Effectiveness of the proposed SgCL loss: As shown in Table 6, the results of No-SgCL are lower than those of the proposed method by 14.21, 15.23, and 30.07 percent in ACC term. Since our consensus representation of multiple views is enhanced by global structure relationship of all samples, and moreover the contrastive learning maximizes the similarity of the consensus representation and view-specific representation from the same sample, minimize the similarity of the representations from different samples with low structure relationship, which improves the clustering performance. More analysis results for SgCL are shown in the supplementary material.

In addition, to further verify the effectiveness of the proposed GCFAggMVC, we visualize different feature representations on different layers after convergence by the t-SNE method [41] in Figure 4. It can be clearly observed that the clustering structure of our learnt consensus features **Ĥ** is clearer than that of the concatenation of view-specific representation **Z** and view-specific representation **H**, and yields well-separated clusters.

5. Conclusion

In this paper, we propose a novel multi-view representation learning network for clustering. We first learn the view-specific features to reconstruct the original data by leveraging the autoencoder model. And then, we design a GCFAgg module, which can learn a global similarity relationship among samples, and moreover enhance a consensus representation by all samples' representation, which makes the representations of these samples with high structure relationship be more similar. Furthermore, we design the SgCL module, which addresses the problem that the representations from different samples in the same cluster is inconsistent in previous contrastive learning methods. Extensive experimental results show that our approach has SOTA performance in both complete and incomplete MVC tasks.

Acknowledgments

References

[1] Mahdi Abavisani and Vishal M Patel. Deep multimodal subspace clustering networks. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1601–1614, 2018. 1

[2] Christian Bauckhage. K-means clustering is matrix factorization. *arXiv preprint arXiv:1512.07548*, 2015. 5

[3] Maria Brbić, Matija Piškorec, Vedrana Vidulin, Anita Kriško, Tomislav Šmuc, and Fran Supek. The landscape of microbial phenotypic traits and associated genes. *Nucleic acids research*, page gkw964, 2016. 5

[4] Xiao Cai, Hua Wang, Heng Huang, and Chris Ding. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12):i16–i24, 2012. 5

[5] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–594, 2015. 2

[6] Guoqing Chao, Shiliang Sun, and Jinbo Bi. A survey on multiview clustering. *IEEE transactions on artificial intelligence*, 2(2):146–168, 2021. 1

[7] Mansheng Chen, Jiaqi Lin, Xianglong Li, Baoyu Liu, Changdong Wang, Dong Huang, and Jianhuang Lai. Representation learning in multi-view clustering: A literature review. *Data Science and Engineering*, 7:225–241, 2022. 1, 2, 5

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 4

[9] Guowang Du, Lihua Zhou, Yudi Yang, Kevin Lü, and Lizhen Wang. Deep multiple auto-encoder-based multi-view clustering. *Data Science and Engineering*, 6(3):323–338, 2021. 1

[10] S El Hajjar, F Dornaika, F Abdallah, and N Barrena. Consensus graph and spectral representation for one-step multi-view kernel based clustering. *Knowledge-Based Systems*, page 108250, 2022. 2

[11] Ruiyi Fang, Liangjian Wen, Zhao Kang, and Jianzhuang Liu. Structure-preserving graph representation learning. *IEEE International Conference on Data Mining*, 2022. 2

[12] 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 *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 5

[13] Hongchang Gao, Feiping Nie, Xuelong Li, and Heng Huang. Multi-view subspace clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 4238–4246, 2015. 2

[14] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE, 2007. 1

[15] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 3

[16] Zhanxuan Hu, Feiping Nie, Wei Chang, Shuzheng Hao, Rong Wang, and Xuelong Li. Multi-view spectral clustering via sparse graph learning. *Neurocomputing*, 384:1–10, 2020. 2

[17] Dong Huang, Chang-Dong Wang, and Jian-Huang Lai. Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity. *arXiv preprint arXiv:2203.11572*, 2022. 5, 6

[18] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, pages 1–8, 2011. 5

[19] Peiguang Jing, Yuting Su, Zhengnan Li, and Liqiang Nie. Learning robust affinity graph representation for multi-view clustering. *Information Sciences*, 544:155–167, 2021. 2

[20] Zhao Kang, Wangtao Zhou, Zhitong Zhao, Junming Shao, Meng Han, and Zenglin Xu. Large-scale multi-view subspace clustering in linear time. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4412–4419, 2020. 2, 5, 6

[21] Guanzhou Ke, Zhiyong Hong, Zhiqiang Zeng, Zeyi Liu, Yangjie Sun, and Yannan Xie. Conan: Contrastive fusion networks for multi-view clustering. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 653–660. IEEE, 2021. 3, 5, 6, 7

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. *Advances in neural information processing systems*, 24, 2011. 5

[24] Yongzhen Li and Husheng Liao. Multi-view clustering via adversarial view embedding and adaptive view fusion. *Applied Intelligence*, 51(3):1201–1212, 2021. 3

[25] Zhenglai Li, Chang Tang, Xinwang Liu, Xiao Zheng, Guanghui Yue, Wei Zhang, and En Zhu. Consensus graph learning for multi-view clustering. *IEEE Transactions on Multimedia*, 2021. 2

[26] Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, Zhaohua Yang, et al. Deep adversarial multi-view clustering network. In *IJCAI*, pages 2952–2958, 2019. 3

[27] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11174–11183, 2021. 3, 5, 7

[28] Zhiping Lin, Zhao Kang, Lizong Zhang, and Ling Tian. Multi-view attributed graph clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 2

[29] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Xifeng Guo, Marius Kloft, and Liangzhong He. Multiview subspace clustering via co-training robust data representation. *IEEE*

Transactions on Neural Networks and Learning Systems, 33(10):5177–5189, 2022. 2

[30] Juncheng Lv, Zhao Kang, Boyu Wang, Luping Ji, and Zenglin Xu. Multi-view subspace clustering via partition fusion. *Information Sciences*, 560:410–423, 2021. 2

[31] David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003. 5

[32] Feiping Nie, Jing Li, Xuelong Li, et al. Self-weighted multiview clustering with multiple graphs. In *IJCAI*, pages 2564–2570, 2017. 2

[33] Erlin Pan and Zhao Kang. Multi-view contrastive graph clustering. *Advances in neural information processing systems*, 34:2148–2159, 2021. 2, 3

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[35] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *International conference on machine learning*, pages 5092–5101, 2019. 5

[36] Jingkuan Song, Hanwang Zhang, Xiangpeng Li, Lianli Gao, Meng Wang, and Richang Hong. Self-supervised video hashing with hierarchical binary auto-encoder. *IEEE Transactions on Image Processing*, 27(7):3210–3221, 2018. 3

[37] Mengjing Sun, Pei Zhang, Siwei Wang, Sihang Zhou, Wenxuan Tu, Xinwang Liu, En Zhu, and Changjian Wang. Scalable multi-view subspace clustering with unified anchors. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3528–3536, 2021. 2, 5, 6

[38] Huayi Tang and Yong Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *Proceedings of the 39th International Conference on Machine Learning*, pages 162:21090–21110, 2022. 1, 2, 5, 6, 7

[39] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 3

[40] 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. 1, 3, 5, 6, 7

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[43] Di Wang, Songwei Han, Quan Wang, Lihuo He, Yumin Tian, and Xinbo Gao. Pseudo-label guided collective matrix factorization for multiview clustering. *IEEE Transactions on Cybernetics*, 52(09):8681–8691, 2022. 5, 6

[44] Qianqian Wang, Zhiqiang Tao, Wei Xia, Quanxue Gao, Xiaochun Cao, and Licheng Jiao. Adversarial multiview clustering networks with adaptive fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3

[45] Xiaobo Wang, Xiaojie Guo, Zhen Lei, Changqing Zhang, and Stan Z Li. Exclusivity-consistency regularized multiview subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 923–931, 2017. 2

[46] Jie Wen, Zheng Zhang, Yong Xu, Bob Zhang, Lunke Fei, and Guo-Sen Xie. Cdimc-net: Cognitive deep incomplete multiview clustering network. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3230–3236, 2020. 5, 7

[47] Jie Wen, Zheng Zhang, Zhao Zhang, Lunke Fei, and Meng Wang. Generalized incomplete multiview clustering with flexible locality structure diffusion. *IEEE transactions on cybernetics*, 51(1):101–114, 2020. 5

[48] H Xiao, K Rasul, and R Vollgraf. A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 5

[49] Jie Xu, Chao Li, Yazhou Ren, Liang Peng, Yujie Mo, Xiaoshuang Shi, and Xiaofeng Zhu. Deep incomplete multiview clustering via mining cluster complementarity. In *Thirty-Six AAAI conference on artificial intelligence*, pages 8761–8769, 2022. 1, 5, 7

[50] Jie Xu, Yazhou Ren, Guofeng Li, Lili Pan, Ce Zhu, and Zenglin Xu. Deep embedded multi-view clustering with collaborative training. *Information Sciences*, 573:279–290, 2021. 1, 5, 6, 7

[51] 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. 1, 2, 3, 5, 6, 7

[52] 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. 1, 3