Deep Multi-View Subspace Clustering with Anchor Graph

Chenhang Cui[1], Yazhou Ren[1,2*], Jingyu Pu[1], Xiaorong Pu[1,2], Lifang He[3]
[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
[2] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
[3] Department of Computer Science and Engineering, Lehigh University, Bethlehem, USA
osallymalone@gmail.com, yazhou.ren@uestc.edu.cn, pujingyu0105@163.com, puxiaor@uestc.edu.cn, lih319@lehigh.edu

Abstract

Deep multi-view subspace clustering (DMVSC) has recently attracted increasing attention due to its promising performance. However, existing DMVSC methods still have two issues: (1) they mainly focus on using autoencoders to nonlinearly embed the data, while the embedding may be suboptimal for clustering because the clustering objective is rarely considered in autoencoders, and (2) they typically have a quadratic or even cubic complexity, which makes it challenging to deal with large-scale data. To address these issues, in this paper we propose a novel deep multi-view subspace clustering method with anchor graph (DMCAG). To be specific, DMCAG firstly learns the embedded features for each view independently, which are used to obtain the subspace representations. To significantly reduce the complexity, we construct an anchor graph with small size for each view. Then, spectral clustering is performed on an integrated anchor graph to obtain pseudo-labels. To overcome the negative impact caused by suboptimal embedded features, we use pseudo-labels to refine the embedding process to make it more suitable for the clustering task. Pseudo-labels and embedded features are updated alternately. Furthermore, we design a strategy to keep the consistency of the labels based on contrastive learning to enhance the clustering performance. Empirical studies on real-world datasets show that our method achieves superior clustering performance over other state-of-the-art methods.

1 Introduction

Subspace clustering has been studied extensively over the years, which assumes that the data points are drawn from low-dimensional subspaces, and could be expressed as a linear combination of other data points. Especially, sparse subspace clustering (SSC) [Elhamifar and Vidal, 2013] has shown the ability to find a sparse representation corresponding to the points from the same subspace. After obtaining the representation of the subspace, the spectral clustering is then applied to obtain the final clustering results. On the other hand, low-rank subspace segmentation was proposed in [Liu *et al.*, 2012] to find a low-rank subspace representation. Despite some state-of-the-art performances have been achieved, most existing methods only focus on single-view clustering tasks.

In many real-world applications, with the exponential growth of data, the description of data has gradually evolved from a single source to multiple sources. For example, a video consists of text, images, and audio. A piece of text can be translated into various languages, and scenes can also be described from different perspectives. These different views often contain complementary information to each other. Making full use of the complementary and consistent information among multiple views could potentially improve the clustering performance.

Considering the diversity of information that comes with multi-view data, the research of multi-view subspace clustering (MVSC) has attracted increasing attention recently. MVSC aims to seek a unified subspace from learning the fusion representation of multi-view data, and then separates data in the corresponding subspace. In the literature, many MVSC methods have been proposed [Zhang *et al.*, 2015; Luo *et al.*, 2018; Li *et al.*, 2019; Wang *et al.*, 2019; Zheng *et al.*, 2020; Liu *et al.*, 2021; Si *et al.*, 2022]. However, one major weakness of existing approaches is their high time and space complexities, which are often quadratic or cubic in the number of samples n. Recently, a number of anchor-based multi-view subspace clustering methods [Chen and Cai, 2011; Sun *et al.*, 2021; Kang *et al.*, 2020; Wang *et al.*, 2022; Liu *et al.*, 2022] have been developed, which can achieve promising performance with a large reduction in storage and computational time. Generally, the anchor graphs are equally weighted and fused into the consensus graph, and then spectral clustering is performed to obtain the clustering result.

On the other hand, inspired by deep neural networks (DNNs), many deep multi-view subspace clustering (DMVSC) methods have been proposed [Peng *et al.*, 2020; Wang *et al.*, 2020; Kheirandishfard *et al.*, 2020; Sun *et al.*, 2019; Zhu *et al.*, 2019; Ji *et al.*, 2017]. However, most DMVSC methods only consider the feature learning ability in networks, their performance is still limited because this learning process is typically independent of the clustering task.

*Corresponding author.

To address the above issues, this paper proposes deep multi-view subspace clustering with anchor graph (DMCAG). DMCAG firstly utilizes deep autoencoders to learn low-dimensional embedded features by optimizing the reconstruction loss for each view independently. For each view, a set of points are chosen by performing k-means on the learned features to construct anchor graphs. Then, we utilize anchor graphs and embedded features as input to get the subspace representation respectively. Once the desired subspace representation is obtained, the clustering result can be calculated by applying the standard spectral clustering algorithm. Unlike the most existing DMVSC methods, the proposed method does not output the clustering result from spectral clustering directly. Instead, we obtain a unified target distribution from this clustering result primarily, which is more robust than that generated by k-means [Xie *et al.*, 2016; Xu *et al.*, 2022], especially for clusters that do not form convex regions or that are not clearly separated. In a self-supervised manner, the Kullback-Leibler (KL) divergence between the unified target distribution and each view's cluster assignments is optimized. We iteratively refine embedding with pseudo-labels derived from the spectral clustering, which in turn help to obtain complementary information and a more accurate target distribution. Besides, to ensure the consistency among different views and avoid affecting the quality of reconstruction, we adopt contrastive learning on the labels instead of latent features. The main contributions of this paper are summarized as follows:

- We propose a novel deep self-supervised model for MVSC. A unified target distribution is generated via spectral clustering which is more robust and can accurately guide the feature learning process. The target distribution and learned features are updated iteratively.
- To boost the model efficiency, we use anchor graph to construct the graph matrix, avoiding constructing a $n \times n$ graph. This strategy can significantly reduce time complexity by sampling anchor points.
- We utilize contrastive learning on pseudo-labels to alleviate the conflict between the consistency objective and the reconstruction objective, thus consistent soft cluster assignments can be obtained among multiple views.
- Extensive experiments on real-world data sets validate the effectiveness and efficiency of the proposed model.

2 Related Work

2.1 Deep Embedded Multi-View Clustering

In recent years, the application of deep learning technology in multi-view clustering has been a hot topic. Deep embedded clustering (DEC) [Xie *et al.*, 2016] utilizes the autoencoder to extract the low-dimensional latent representation from raw features and then optimizes the student's t-distribution and target distribution of the feature representation to achieve clustering. In contrast, traditional multi-view clustering algorithms mostly use linear and shallow embedding to learn the latent structure of multi-view data.

However, these methods cannot utilize the nonlinear property of data availably, which is crucial to reveal a complex clustering structure [Ren *et al.*, 2022]. Deep embedded multi-view clustering with collaborative training (DEMVC) [Xu *et al.*, 2021a] is a novel framework for multi-view clustering, in which a shared scheme of the auxiliary distribution is used to improve the performance of the clustering. By assuming that clustering structures with high discriminability play a significant role in clustering, self-supervised discriminative feature learning for multi-view clustering (SDMVC) [Xu *et al.*, 2022] leverages the global discriminative information contained in all views' embedded features. During the process, the global information will guide the feature learning process of each view. The autoencoder is usually utilized to capture the most important features present in the data. A suitable autoencoder can obtain more robust representations. Deep embedding clustering based on contractive autoencoder (DECCA) [Diallo *et al.*, 2021] simultaneously disentangles the problem of learned representation by preserving important information from the initial data while pushing the original samples and their augmentations together. With the introduction of the contractive autoencoders, the learned features are better than normal autoencoders. The available information provided by latent graph is often ignored, deep embedded multi-view clustering via jointly learning latent representations and graphs (DMVCJ) [Huang *et al.*, 2022] utilizes the graphs from latent features to promote the performance of deep MVC models. By learning the latent graphs and feature representations jointly, the available information implied in latent graph can improve the clustering performance. The self-supervised manner is widely used in deep clustering, which is valuable to migrate it to other clustering algorithms.

2.2 Multi-View Subspace Clustering

Although many methods exist in subspace clustering, such as low-rank representation subspace clustering (LRR) [Liu *et al.*, 2012], sparse subspace clustering (SSC) [Elhamifar and Vidal, 2013], most of the multi-view subspace clustering methods adopt self-representation to obtain the subspace representation. Low-rank tensor constrained multi-view subspace clustering (LMSC) [Zhang *et al.*, 2017] learns the latent representation based on multi-view features, and generates a common subspace representation rather than that of individual view. Flexible multi-view representation learning for subspace clustering (FMR) [Li *et al.*, 2019] avoids using partial information for data reconstruction and makes the latent representation well adapted to subspace clustering.

Most of existing MVSC methods are challenging to apply in large-scale data sets. Fortunately, inspired by the idea of anchor graph which can help reduce both storage and computational time, a lot of anchor-based MVSC methods have been proposed. Large-scale multi-view subspace clustering (LMVSC) [Kang *et al.*, 2020] can be solved in linear time, which selects a small number of instances to construct anchor graphs for each view, and then integrates all anchor graphs from each views. It thus can perform spectral clustering on a small graph. Efficient one-pass multi-view subspace clustering with consensus anchor (CGMSC) [Liu *et al.*, 2022] unifies fused graph construction and anchor learning into a unified and flexible framework so that they seamlessly contribute mutually and boost performance. Fast multi-view

anchor correspondence clustering (FMVACC) [Wang *et al.*, 2022] finds that the selected anchor sets in multi-view data are not aligned, which may lead to inaccurate graph fusion and degrade the clustering performance, so an anchor alignment module is proposed to solve the anchor-unaligned problem (AUP).

With the development of deep neural networks, a number of deep multi-view subspace clustering methods have been proposed. Deep subspace clustering with ℓ_1-norm (DSC-L1) [Peng *et al.*, 2020] learns nonlinear mapping functions for data to map the original features into another space, and then the affinity matrix is calculated in the new space. Deep multi-view subspace clustering with unified and discriminative learning (DMSC-UDL) [Wang *et al.*, 2020] integrates local and global structure learning simultaneously, which can make full use of all the information of the original multi-view data. Different from the above-mentioned deep multi-view clustering methods, we combine representation learning and spectral clustering into a unified optimization framework, the clustering results can be sufficiently exploited to guide the representation learning for each view.

3 Methodology

Problem Statement. Given multi-view data $X = \{X^v \in \mathbb{R}^{d_v \times n}\}_{v=1}^{V}$ with V views, d_v is the dimension of the v-th view, and n is the instance number. The target of MVSC is to divide the given instances into k clusters.

3.1 Motivation

Subspace clustering aims to find out an underlying subspace which expresses each point as a linear combination of other points. The final clustering assignment is obtained by performing spectral clustering on the learned subspace. Basically, it can be mathematically formulated as:

$$\begin{aligned} \min_{X^v} \sum_{v=1}^{V} \|X^v - X^v S^v\|_F^2 + \gamma \|S^v\|_F^2 \\ s.t.\ \ S^v \geqslant \mathbf{0}, S^v \mathbf{1} = \mathbf{1}, \end{aligned} \tag{1}$$

where $S^v \in R^{n \times n}$ is the learned subspace of v-th view, $\mathbf{1} \in R^{n \times 1}$ is a vector where all its elements are equal to one, and γ is a trade-off coefficient that controls the sparsity of S^v. The constraints on S^v ensure that S^v is non-negative and $\sum_j S_{ij} = 1$. When the adjacency graphs are obtained, spectral clustering can be performed on S^v to get the clustering result. Existing MVSC methods aim at learning concrete subspace effectively and mining global information of all views to improve the clustering assignment quality, but still meet some challenges:

(1) Most MVSC methods require at least $O(n^2 k)$ time complexity to calculate the clustering result on raw features, which requires more storage and time and is also difficult to deal with large-scale datasets. Additionally, some DMVSC methods such as [Ji *et al.*, 2017; Li *et al.*, 2021] focus on learning subspace by nonlinearly embedding the data. Since the connection between the learned subspace and clustering is weak, the learned features may not be optimal for the clustering task.

(2) From existing self-supervised MVC methods, we can observe that most of them heavily depend on the quality of latent representations to supervise the learning process. [Xu *et al.*, 2022] fuses the embedded features of all views and utilizes k-means [MacQueen, 1967] to obtain the global target distribution. However, k-means does not perform well on some data structure such as non-convex structure compared with spectral clustering [Ng *et al.*, 2001] which is more robust to different distributions. Hence, their obtained pseudo-labels may not reflect clear clustering structure to guide the latent feature learning process.

(3) Some views of an instance might have wrong clustering assignment, leading to the clustering inconsistency. Some deep-learning based MVC methods achieve the consistency by directly learning common information on latent features [Cheng *et al.*, 2021], which may reduce the complementarity of each view's information due to the conflict between learning exact latent representations and achieving consistency.

To address the above-mentioned issues, we propose a novel DMCAG framework as shown in Fig. 1. We use the anchor method to construct the graph matrix on latent features of each view which requires less time and storage. After that, we obtain global pseudo-labels by performing spectral clustering on the integrated anchor graph. Since spectral clustering is more robust to the data distribution, our self-supervised process can learn higher discriminative information from each view. Then, we adopt contrastive learning on pseudo-labels of latent features to maintain the view-private information and achieve the clustering consistency among all views.

3.2 Learning Anchor Graph via Autoencoders

As for much redundancy in the raw data, we utilize the deep autoencoder to extract the latent representations of all views. Through the encoder $f_{\theta^v}^v$ and decoder $g_{\phi^v}^v$, where θ^v and ϕ^v are learnable parameters, X^v is encoded as $Z^v \in R^{l \times n}$ (l is the same for all views) via $f_{\theta^v}^v$ and Z^v is decoded as $\hat{X}^v$ via $g_{\phi^v}^v$. The reconstruction loss is defined as:

$$L_r = \sum_{v=1}^{V} L_r^v = \sum_{v=1}^{V} \|X^v - g_{\phi^v}^v(f_{\theta^v}^v(X^v))\|_F^2. \tag{2}$$

Inspired by [Kang *et al.*, 2020], we adopt the anchor graph to replace the full adjacency matrix S, which is formulated as:

$$\begin{aligned} \min_{C^v} L_a^v = \sum_{v=1}^{V} \|Z^v - A^v (C^v)^T\|_F^2 + \gamma \|C^v\|_F^2 \\ s.t.\ \ C^v \geqslant 0, (C^v)^T \mathbf{1} = \mathbf{1}, \end{aligned} \tag{3}$$

where $A^v \in R^{l \times m}$ (m is the anchor number) is a set of clustering centroids through k-means on the embedded features Z^v, and $C^v \in R^{n \times m}$ is the anchor graph matrix reflecting relationship between Z^v and A^v. The problem above can be solved by convex quadratic programming. We refer the readers to [Wolfe, 1959] for more details about quadratic programming.

3.3 Spectral Self-Supervised Learning

As stated in [Ng *et al.*, 2001], for clusters that are not clearly separated or do not form convex regions, the spectral method

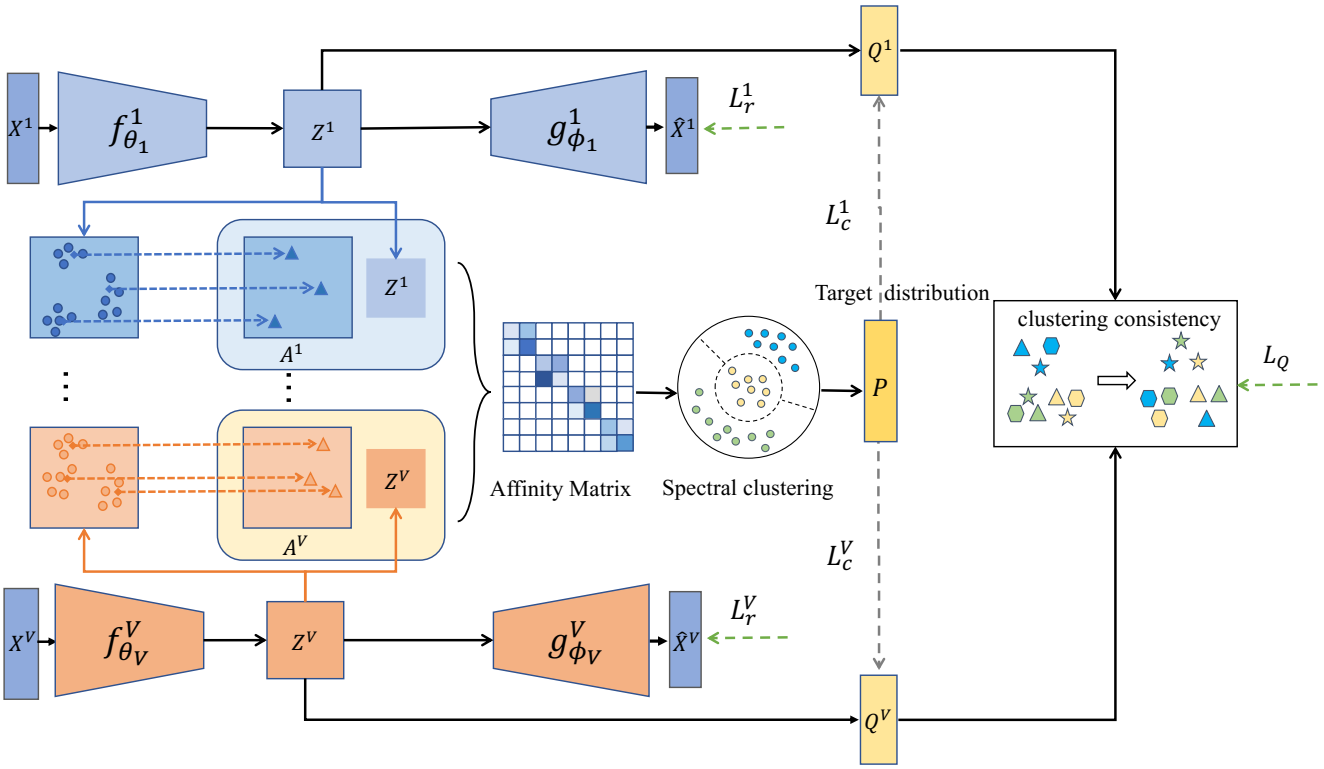


Figure 1: The framework of DMCAG. For the v-th view, X^v denotes the input data, Z^v denotes the embedded features, A^v is a set of clustering centroids with the number of m, and Q^v is the soft cluster assignment distribution. P denotes the unified target distribution obtained through spectral clustering.

can also reliably find clustering assignment. Hence, we use spectral clustering to obtain more robust global target distribution to guide the self-training. Spectral clustering [Ng *et al.*, 2001] can be mathematically described as finding $Q \in R^{n\times k}$ by maximizing:

$$\max_{Q} Tr\,(Q^T S\, Q) \;\; s.t.\; Q^T Q = I. \tag{4}$$

Following Theorem 1 introduced by [Chen and Cai, 2011; Kang *et al.*, 2020], we present an approach to approximate the singular vectors of S in latent space.

Theorem 1. *[Chen and Cai, 2011; Kang* et al.*, 2020] Given a similarity matrix S, which can be decomposed as $(C^T)C$. Define singular value decomposition (SVD) of C as $U\wedge V^T$, then we have*

$$\max_{Q^TQ=I} Tr\,(Q^T S\, Q) \Longleftrightarrow \min_{Q^TQ=I,H} \|C - QH^T\|_F^2. \tag{5}$$

And the optimal solution Q^* is equal to U.

Proof. From Eq. (5), one can observe that the optimal $H^* = C^T Q$. Substituting $H^* = C^T Q$ into Eq. (5), the following equivalences hold

$$\min_{Q^TQ=I,H} \|C - QH^T\|_F^2 \Longleftrightarrow \min_{Q^TQ=I} \|C - QQ^TC\|_F^2$$
$$\Longleftrightarrow \max_{Q^TQ=I} Tr\,(Q^T CC^T Q) \Longleftrightarrow \max_{Q^TQ=I} Tr\,(Q^T S\, Q).$$

Furthermore, one can obtain

$$S = CC^T = U\wedge V^T (U\wedge V^T)^T = U\wedge (V^T V)\wedge U^T = U\wedge^2 U^T.$$

Therefore, we could use left singular vectors of C to approximate the eigenvectors of S. □

According to Theorem 1, we calculate eigenvectors $\{U^v\}_{v=1}^V$ of $\{C^v\}_{v=1}^V$ to approximate the eigenvectors of full similarity matrix. To fully exploit the complementary information across all views, we concatenate all eigenvectors $U = \{U^1, U^2, \ldots, U^V\} \in R^{n\times(Vk)}$ to generate the global feature via the spectral method. After obtaining the global feature U, we apply k-means to calculate the cluster centroids $\{\mu_j\}_{j=1}^k$:

$$\min_{\mu_1,\ldots,\mu_k} \sum_{i=1}^{n}\sum_{j=1}^{k} \left\|U(i,:) - \mu_j\right\|^2. \tag{6}$$

Similiar to DEC [Xie *et al.*, 2016], which is a popular single-view deep clustering method utilizing Student's t-distribution [Van der Maaten and Hinton, 2008], the soft clustering assignment t_{ij} between global feature U and each cluster centroid μ_j could be computed as:

$$t_{ij} = \frac{(\alpha + \left\|U(i,:) - \mu_j\right\|^2)^{-1}}{\sum_j (\alpha + \left\|U(i,:) - \mu_j\right\|^2)^{-1}}. \tag{7}$$

To increase the discriminability of the global soft assignments, the target distribution P is formulated as:

$$p_{ij} = \frac{(t_{ij}^2 / \sum_i t_{ij})}{\sum_j (t_{ij}^2 / \sum_i t_{ij})}. \tag{8}$$

We obtain soft clustering assignment (pseudo-label) $Q^v = \left[q_1^v, q_2^v, \ldots, q_N^v\right]$ of each view, where q_{ij}^v can be considered as the probability of the i-th instance belonging to the j-th cluster in the v-th view. It is defined as:

$$q_{ij}^v = \frac{(1 + \left\| z_i^v - \mu_j^v \right\|^2)^{-1}}{\sum_j (1 + \left\| z_i^v - \mu_j^v \right\|^2)^{-1}}, \tag{9}$$

where μ_j^v denotes the j-th cluster centroid of the v-th view.

Overall, we use Kullback-Leibler divergence between the unified target distribution P and view-private soft assignment distribution Q^v to guide autoencoders to optimize latent features containing higher discriminative information, which can be formulated as:

$$L_c^v = D_{KL}(P \| Q^v) = \sum_{v=1}^{V} \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij} \log \frac{p_{ij}}{q_{ij}^v}. \tag{10}$$

As the target distribution obtained from spectral clustering is adaptive to different data distributions, we can get more explicit cluster structures to guide the self-training process compared with k-means clustering. To extract embedded features that reflect correct information of raw features and learn an accurate assignment for clustering, we jointly optimize the reconstruction of autoencoders and self-supervised learning. The total loss function L_s is defined as:

$$L_s = \sum_{v=1}^{V} (L_r^v + L_c^v). \tag{11}$$

3.4 Label Consistency Learning

To guarantee the same soft assignment distribution of all views represent the same cluster, we need to achieve the consistency of pseudo-labels. We adopt contrastive learning to the soft assignment obtained from Eq. (9). For the m-th view, $Q^m(:, j)$ have $(Vk - 1)$ pairs, where the $(V - 1)$ pairs $\{Q^m(:, j), Q^n(:, j)\}_{m \neq n}$ are positive and the rest $V(k - 1)$ pairs are negative. Thereby the contrastive loss can be defined as:

$$L_Q^{mn} = -\frac{1}{k} \sum_{j=1}^{k} \log \frac{e^{d(Q^m(:,j), Q^n(:,j))/\tau}}{\sum_{k'=1}^{k} \sum_{v=m,n} e^{d(Q^m(:,j), Q^v(:,k'))/\tau} - e^{1/\tau}}, \tag{12}$$

where $d(\cdot, \cdot)$ represents the cosine distance to meausure the similiarty between two labels, τ is the temperature parameter. Moreover, to avoid the samples being assigned into a single cluster, we use the cross entropy as a regularization term. Generally, the label consistency learning is formulated as:

$$L_Q = \frac{1}{2} \sum_{m=1}^{V} \sum_{n \neq m} L_Q^{mn} + \sum_{m=1}^{V} \sum_{j=1}^{k} s_j^m \log s_j^m, \tag{13}$$

Algorithm 1 Deep Multi-View Subspace Clustering with Anchor Graph (DMCAG)

Input: multi-view dataset X, cluster number k.
Initialization: Get $\{\theta^v, \phi^v, \mu^v, A^v\}_{v=1}^V$ by pretraining autoencoders and k-means. Initialize $\{C^v\}_{v=1}^V$ via quadratic programming.
Update $\{\theta^v, \phi^v\, Q^v\}_{v=1}^V$ by performing self-supervised learning via Eq. (11).
Performing contrastive learning on $\{Q^v\}_{v=1}^V$ via Eq. (13).
Obtain $\{C^v\}_{v=1}^V$ via Eq. (3).
Output: Cluster assignment y via Eq. (6).

where $s_j^m = \frac{1}{N} \sum_{i=1}^{N} q_{ij}^m$. After finetuning the labels via contrastive learning, the similarities of positive pairs are enhanced and thereby the obtained latent features have clearer clustering structure. At last, the clustering prediction y is obtained through the target distribution P calculated by performing k-means on U.

3.5 Optimization

The detailed optimization procedure is summarized in Algorithm 1. We adopt the Adam method to train the autoencoders. At the beginning, autoencoders are initialized by Eq. (2). After that, we solve Eq. (3) via convex quadratic programming to obtain U and calculate the global target distribution P. Then, the spectral self-supervised learning is adopted to learn more representive embeddings. After performing the self-supervised learning, the contrastive learning is conducted to achieve the clustering consistency. At last, we run k-means on U to obtain the final clustering result y:

$$y_i = \underset{j}{\operatorname{argmax}} \, (p_{ij}). \tag{14}$$

4 Experiments

4.1 Experimental Settings

Dataset	Sample	View	Dimension
MNIST-USPS	5000	2	[[28,28], [28,28]]
Multi-COIL-10	720	3	[[32,32], [32,32], [32, 32]]
BDGP	2500	2	[1750, 79]
UCI-digits	2000	3	[240, 76, 64]
Fashion-MV	10000	3	[784, 784, 784]
HW	2000	6	[216, 76, 64, 6, 240, 47]

Table 1: The statistics of experimental datasets.

Datasets. As shown in Table 1, our experiments are carried out on six datasets. Specifically, **MNIST-USPS** [Peng *et al.*, 2019] collects from two handwritten digital image datasets, which are treated as two views. **Multi-COIL-10** [Xu *et al.*, 2021b] collects 720 grayscale object images with size of 32×32 from 10 clusters, where the different views represent various poses of objects. **BDGP** [Cai *et al.*, 2012] **UCI-digits**[1] is a collection of 2000 samples with 3 views,

[1]https://archive.ics.uci.edu/ml/datasets/Multiple%2BFeatures

Datasets	K-means	SC	DEC	CSMSC	FMR	SAMVC	LMVSC	CGMSC	FMVACC	Ours
					ACC					
MNIST-USPS	76.78	65.96	73.10	72.68	63.02	69.65	38.54	91.22	98.67	**99.58**
Multi-COIL-10	73.36	33.75	74.01	97.64	78.06	84.31	63.79	99.86	93.20	**100.00**
BDGP	43.24	51.72	94.78	53.48	95.12	51.31	35.85	45.60	58.63	**98.00**
UCI-digits	79.50	63.35	87.35	88.20	80.10	74.20	74.60	76.95	89.47	**95.60**
Fashion-MV	70.93	53.54	67.07	77.61	-	62.86	43.43	82.79	79.62	**97.89**
HW	75.45	77.69	81.13	89.80	86.05	76.37	91.65	69.10	89.45	**97.90**
					NMI					
MNIST-USPS	72.33	58.11	71.46	72.64	60.90	60.99	63.09	88.24	96.74	**99.86**
Multi-COIL-10	76.91	12.31	77.43	96.17	80.04	92.09	75.83	99.68	93.39	**100.00**
BDGP	56.94	58.91	86.92	39.92	87.69	45.15	36.69	27.15	36.81	**94.69**
UCI-digits	77.30	66.60	79.50	80.68	72.13	74.73	74.93	77.62	84.35	**91.10**
Fashion-MV	65.61	57.72	72.34	77.91	-	68.78	46.02	86.48	76.31	**95.40**
HW	78.58	86.91	82.61	82.95	76.49	84.41	84.43	81.83	85.98	**95.25**
					ARI					
MNIST-USPS	63.53	48.64	63.23	64.08	49.73	74.58	27.74	86.46	97.65	**99.07**
Multi-COIL-10	64.85	14.02	65.66	94.89	70.59	88.75	55.05	99.69	92.47	**100.00**
BDGP	26.04	31.56	87.02	33.59	88.38	19.60	35.06	21.99	44.25	**95.11**
UCI-digits	71.02	54.07	75.69	76.72	65.53	74.25	74.27	70.08	83.33	**90.59**
Fashion-MV	56.89	42.61	62.91	70.28	-	56.65	41.12	78.27	72.92	**95.48**
HW	66.72	75.26	74.25	79.50	72.59	73.87	83.20	69.54	85.04	**95.40**

Table 2: Results of all methods on six datasets. The best result in each row is shown in bold and the second-best is underlined.

	Components			MNIST-USPS			Fashion-MV			BDGP		
	L_r	L_s	L_Q	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
(A)	✓	✗	✗	64.26	65.37	52.03	62.80	71.57	55.56	37.96	36.42	14.10
(B)	✓	✓	✗	83.44	82.34	73.67	75.12	79.65	67.08	61.88	65.13	46.56
(C)	✓	✗	✓	84.10	90.21	82.79	93.38	90.78	87.55	71.28	64.34	56.15
(D)	✓	✓	✓	99.58	98.86	99.07	97.89	95.40	95.48	98.00	94.69	95.11

Table 3: Ablation studies on DMCAG.

which has 10 categories. **Fashion-MV** [Xiao *et al.*, 2017] contains images from 10 categories, where we treat three different styles of one object as three views. **Handwritten Numerals (HW)**[2] contains 2000 samples from 10 categories corresponding to numerals 0-9. Each sample has six visual views.

Comparison Methods. Comparsion methods include 3 traditional single-view clustering methods, i.e., k-means [MacQueen, 1967], SC (Spectral clustering [Ng *et al.*, 2001]), and DEC (Deep embedded clustering [Xie *et al.*, 2016]), and 6 state-of-art MVC methods, i.e., CSMSC (Consistent and specific multi-view subspace clustering[Luo *et al.*, 2018]), FMR (Flexible multi-view representation learning for subspace clustering), SAMVC (Self-paced and auto-weighted multi-view clustering [Ren *et al.*, 2020]), LMVSC (Large-scale multi-view subspace clustering in linear time [Kang *et al.*, 2020]), CGMSC (Multi-view subspace clustering with adaptive locally consistent graph regularization [Liu *et al.*, 2021]), and FMVACC (Fast multi-view anchor-correspondence clustering [Wang *et al.*, 2022]).

Evaluation Metrics. We evaluate the effectiveness of clustering by three commonly used metrics, i.e., clustering accuracy (ACC), normalized mutual information (NMI), and adjusted rand index (ARI). A higher value of each evaluation metric indicates a better clustering performance.

Implementation. The convolutional (Conv) and fully connected (Fc) neural networks are applied according to different types of data. For the image datasets, i.e., MNIST-USPS and Multi-COIL-10, we use the convolutional autoencoder (CAE) for each view to learn embedded features. The encoder is Input-Conv_4^{32}-Conv_4^{64}-Conv_4^{64}-Fc_{10}. Since all views of BDGP, UCI-digits, Fashion-MV, and HW are vector data, we utilize fully connected autoencoder. For each view, the encoder is Input-Fc_{500}-Fc_{500}-Fc_{2000}-Fc_{10}. All the decoders are symmetric with the corresponding encoders. Following [Kang *et al.*, 2020], we select anchor numbers in the range [10, 100]. We select γ from {0.1, 1, 10}. Temperature parameter τ is set to 1 and α is set to 0.001 for all experiments. All experiments are performed on Windows PC with Intel (R) Core (TM) i5-12600K CPU@3.69 GHz, 32.0 GB RAM, and GeForce RTX 3070ti GPU (8 GB caches). For fair comparison, all baselines are tuned to the best performance according to the corresponding papers.

[2]https://archive.ics.uci.edu/ml/datasets.php

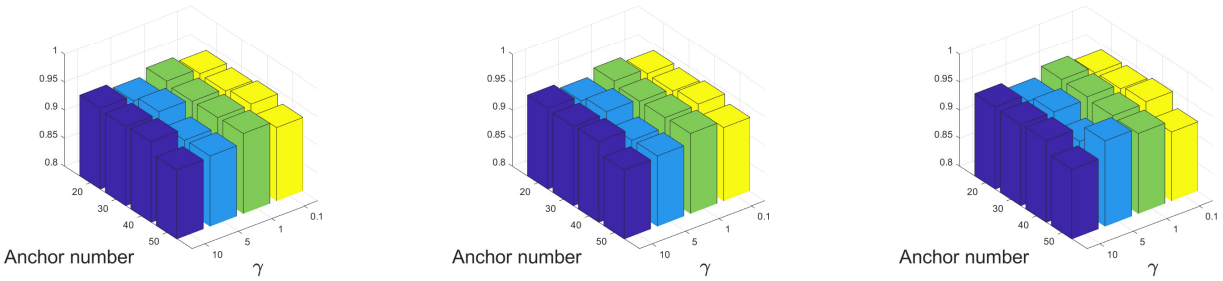
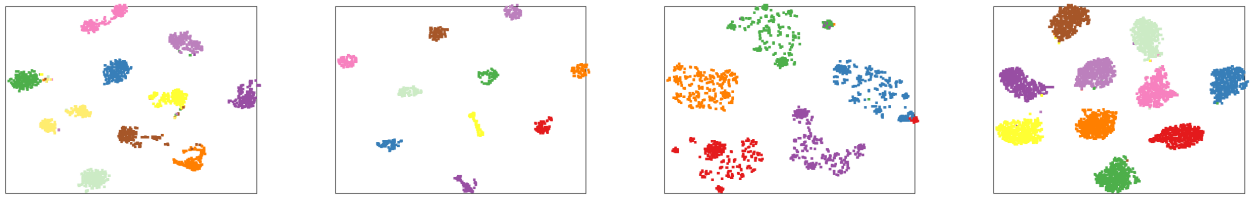


Figure 2: Clustering performance with different parameter settings on HW dataset.



(a) HW (b) Multi-COIL-10 (c) BDGP (d) MNIST-USPS

Figure 3: Visualization of the clustering results on four datasets.

4.2 Results and Analysis

Comparison with Baselines. The comparison of DMCAG and baseline methods is shown in Table 2, where the best result is bolded in each row and the second-best result is underlined. Due to the high complexity, the result of FMR is not obtained on Fasion-MV dataset after running 24 hours. From Table 2, we can observe that the proposed method achieves the best performance among all baseline models on six datasets, illustrating the validity of our method. Especially for Multi-COIL-10 and Fasion-MV, our method makes greater improvement than existing methods. The main reason is that we conduct clearer self-supervised learning with latent anchor graphs and achieve clustering consistency via contrastive learning on pseudo-labels. Besides, many subspace clustering methods do not perform well on BDGP dataset due to the high dimension of one view's feature, however, our method outperforms due to the ability of extracting features from high-dimension data.

Ablation Studies. To verify the effectiveness of the proposed method, we further conduct a set of ablation studies on the loss components from Eqs. (10) and (13). L_r is the reconstruction loss of autoencoders. L_s aims to mine global information and supervise the learning process with global target distribution. L_Q is optimized to achieve the consistency among all views. In the Table 3, (A) is optimized without self-supervised learning and contrastive learning, (B) is optimized without contrastive learning, and (C) is optimized without self-supervised process. As shown in Table 3, we can find that (B) and (C) perform better than (A). (D) (our proposed model with all loss components) performs the best, illustrating that each loss component is important for the final clustering result.

Parameter Sensitivity Analysis. As main hyperparameters of the proposed method are anchor number and sparsity coefficient, we test the general clustering performance with different settings to show the stability of DMCAG. From Fig. 2 we can observe that our method in a certain range of parameters is insensitive to the clustering results. In addition, we can find that too many anchors will reduce clustering performance as the anchors' information may not be representative. A too large or too small γ also reduces the final clustering performence, of which the reason is that the inappropriate sparsity constraint could lead to extra errors.

Visualization of Clustering Results. We visualize the clustering results on four datasets via t-SNE [Van der Maaten and Hinton, 2008], which reduces the dimension of the extracted feature vectors to 2D. As shown in Fig. 3, where different colors denote the labels of different nodes, we can find that the final clustering structures are clearly visible, especially for Multi-COIL-10 and MNIST-USPS, which further demonstrates the effectiveness of our DMCAG method.

5 Conclusion

In this paper, we propose a novel DMCAG framework for deep multi-view subspace clustering. By introducing the latent anchor graph and spectral self-supervised learning, we can effectively perform spectral clustering to obtain more robust global target distribution and significantly improve the latent features structure for clustering. In addition, we adopt the contrastive learning on the soft assignment of each view to achieve the consistency among all views.

Acknowledgments

This work was supported in part by Sichuan Science and Technology Program (Nos. 2022YFS0047 and 2022YFS0055), Medico-Engineering Cooperation Funds from the University of Electronic Science and Technology of China (No. ZYGX2021YGLH022), National Science Foundation (No. MRI 2215789), and Lehigh's grants (Nos. S00010293 and 001250).

References

[Cai *et al.*, 2012] 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.

[Chen and Cai, 2011] Xinlei Chen and Deng Cai. Large scale spectral clustering with landmark-based representation. In *AAAI*, pages 313–318, 2011.

[Cheng *et al.*, 2021] Jiafeng Cheng, Qianqian Wang, Zhiqiang Tao, Deyan Xie, and Quanxue Gao. Multi-view attribute graph convolution networks for clustering. In *IJCAI*, pages 2973–2979, 2021.

[Diallo *et al.*, 2021] Bassoma Diallo, Jie Hu, Tianrui Li, Ghufran Ahmad Khan, Xinyan Liang, and Yimiao Zhao. Deep embedding clustering based on contractive autoencoder. *Neurocomputing*, 433:96–107, 2021.

[Elhamifar and Vidal, 2013] Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *TPAMI*, 35(11):2765–2781, 2013.

[Huang *et al.*, 2022] Zongmo Huang, Yazhou Ren, Xiaorong Pu, and Lifang He. Deep embedded multi-view clustering via jointly learning latent representations and graphs. *arXiv preprint arXiv:2205.03803*, 2022.

[Ji *et al.*, 2017] Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. *NeurIPS*, pages 23–32, 2017.

[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 *AAAI*, pages 4412–4419, 2020.

[Kheirandishfard *et al.*, 2020] Mohsen Kheirandishfard, Fariba Zohrizadeh, and Farhad Kamangar. Multi-level representation learning for deep subspace clustering. In *WACVW*, pages 2039–2048, 2020.

[Li *et al.*, 2019] Ruihuang Li, Changqing Zhang, Qinghua Hu, Pengfei Zhu, and Zheng Wang. Flexible multi-view representation learning for subspace clustering. In *IJCAI*, pages 2916–2922, 2019.

[Li *et al.*, 2021] Kai Li, Hongfu Liu, Yulun Zhang, Kunpeng Li, and Yun Fu. Self-guided deep multiview subspace clustering via consensus affinity regularization. *IEEE T CYBERNETICS*, 2021.

[Liu *et al.*, 2012] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *TPAMI*, 35(1):171–184, 2012.

[Liu *et al.*, 2021] Xiaolan Liu, Gan Pan, and Mengying Xie. Multi-view subspace clustering with adaptive locally consistent graph regularization. *Neural Computing and Applications*, 33(22):15397–15412, 2021.

[Liu *et al.*, 2022] Suyuan Liu, Siwei Wang, Pei Zhang, Kai Xu, Xinwang Liu, Changwang Zhang, and Feng Gao. Efficient one-pass multi-view subspace clustering with consensus anchors. In *AAAI*, pages 7576–7584, 2022.

[Luo *et al.*, 2018] Shirui Luo, Changqing Zhang, Wei Zhang, and Xiaochun Cao. Consistent and specific multi-view subspace clustering. In *AAAI*, pages 3730–3737, 2018.

[MacQueen, 1967] James MacQueen. Classification and analysis of multivariate observations. In *BSMSP*, pages 281–297, 1967.

[Ng *et al.*, 2001] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, pages 1–8, 2001.

[Peng *et al.*, 2019] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *ICML*, pages 5092–5101, 2019.

[Peng *et al.*, 2020] Xi Peng, Jiashi Feng, Joey Tianyi Zhou, Yingjie Lei, and Shuicheng Yan. Deep subspace clustering. *TNNLS*, 31(12):5509–5521, 2020.

[Ren *et al.*, 2020] Yazhou Ren, Shudong Huang, Peng Zhao, Minghao Han, and Zenglin Xu. Self-paced and auto-weighted multi-view clustering. *Neurocomputing*, 383:248–256, 2020.

[Ren *et al.*, 2022] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S. Yu, and Lifang He. Deep clustering: A comprehensive survey. *arXiv preprint arXiv:2210.04142*, 2022.

[Si *et al.*, 2022] Xiaomeng Si, Qiyue Yin, Xiaojie Zhao, and Li Yao. Consistent and diverse multi-view subspace clustering with structure constraint. *Pattern Recognition*, 121:108196, 2022.

[Sun *et al.*, 2019] Xiukun Sun, Miaomiao Cheng, Chen Min, and Liping Jing. Self-supervised deep multi-view subspace clustering. In *ACML*, pages 1001–1016, 2019.

[Sun *et al.*, 2021] 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 *ACM MM*, pages 3528–3536, 2021.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

[Wang *et al.*, 2019] Xiaobo Wang, Zhen Lei, Xiaojie Guo, Changqing Zhang, Hailin Shi, and Stan Z Li. Multi-view subspace clustering with intactness-aware similarity. *Pattern Recognition*, 88:50–63, 2019.

[Wang *et al.*, 2020] Qianqian Wang, Jiafeng Cheng, Quanxue Gao, Guoshuai Zhao, and Licheng Jiao.

Deep multi-view subspace clustering with unified and discriminative learning. *TMM*, 23:3483–3493, 2020.

[Wang *et al.*, 2022] Siwei Wang, Xinwang Liu, Suyuan Liu, Jiaqi Jin, Wenxuan Tu, Xinzhong Zhu, and En Zhu. Align then fusion: Generalized large-scale multi-view clustering with anchor matching correspondences. *arXiv preprint arXiv:2205.15075*, 2022.

[Wolfe, 1959] Philip Wolfe. The simplex method for quadratic programming. *Econometrica: Journal of the Econometric Society*, pages 382–398, 1959.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

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

[Xu *et al.*, 2021a] 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.

[Xu *et al.*, 2021b] 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 *ICCV*, pages 9234–9243, 2021.

[Xu *et al.*, 2022] Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, Philip S. Yu, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *TKDE*, 2022.

[Zhang *et al.*, 2015] Changqing Zhang, Huazhu Fu, Si Liu, Guangcan Liu, and Xiaochun Cao. Low-rank tensor constrained multiview subspace clustering. In *ICCV*, pages 1582–1590, 2015.

[Zhang *et al.*, 2017] Changqing Zhang, Qinghua Hu, Huazhu Fu, Pengfei Zhu, and Xiaochun Cao. Latent multi-view subspace clustering. In *ICCV*, pages 4279–4287, 2017.

[Zheng *et al.*, 2020] Qinghai Zheng, Jihua Zhu, Zhongyu Li, Shanmin Pang, Jun Wang, and Yaochen Li. Feature concatenation multi-view subspace clustering. *Neurocomputing*, 379:89–102, 2020.

[Zhu *et al.*, 2019] Pengfei Zhu, Binyuan Hui, Changqing Zhang, Dawei Du, Longyin Wen, and Qinghua Hu. Multi-view deep subspace clustering networks. *arXiv preprint arXiv:1908.01978*, 2019.