MCoCo: Multi-level Consistency Collaborative Multi-view Clustering

Yiyang Zhou[1], Qinghai Zheng[2], Wenbiao Yan[1], Yifei Wang[1], Pengcheng Shi[1], Jihua Zhu[1,*]

[1]School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China

[2]College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China

Abstract—Multi-view clustering can explore consistent information from different views to guide clustering. Most existing works focus on pursuing shallow consistency in the feature space and integrating the information of multiple views into a unified representation for clustering. These methods did not fully consider and explore the consistency in the semantic space. To address this issue, we proposed a novel Multi-level Consistency Collaborative learning framework (MCoCo) for multi-view clustering. Specifically, MCoCo jointly learns cluster assignments of multiple views in feature space and aligns semantic labels of different views in semantic space by contrastive learning. Further, we designed a multi-level consistency collaboration strategy, which utilizes the consistent information of semantic space as a self-supervised signal to collaborate with the cluster assignments in feature space. Thus, different levels of spaces collaborate with each other while achieving their own consistency goals, which makes MCoCo fully mine the consistent information of different views without fusion. Compared with state-of-the-art methods, extensive experiments demonstrate the effectiveness and superiority of our method. Our code is released on https://github.com/YiyangZhou/MCoCo.

Index Terms—Multi-view clustering, Consistency collaborative, Semantic consensus information.

I. INTRODUCTION

MULTI-VIEW data are collected by different collectors and feature extractors, and different views show heterogeneity. Compared with the traditional single-view data, it is informative and can provide a more comprehensive description of objects [1]–[5]. Thanks to these advantages, multi-view clustering (MVC) has attracted more and more attention in recent years. Existing MVC methods can be roughly divided into traditional methods and deep methods.

Traditional methods can be further divided into three subcategories: (1) Subspace-based clustering methods [2], [6]–[11], where a shared low-dimensional representation that integrates multi-view information and a similarity matrix is mined for clustering. (2) Clustering method based on non-negative matrix decomposition [12]–[14], which decomposes each view into a low-rank matrix for clustering. (3) Graph-based clustering method [15]–[20], mining graph structure information to guide multi-view clustering.

The deep neural network has shown excellent performance in many fields in recent years [21]–[23]. In order to utilize the ability of the deep network to capture nonlinear features and deal with clustering tasks of large-scale data [24], [25], many MVC methods based on the deep network have appeared
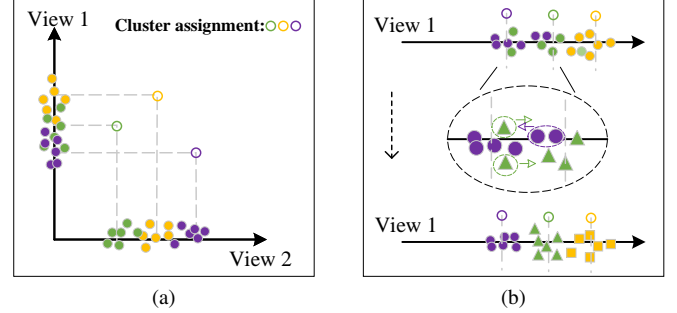
Fig. 1. An illustrative example of our motivation. (a) Showing the mapping between two views. (b) Taking view 1 as an example, it shows the collaborative learning process of semantic labels and cluster assignments. If there are three categories in the dataset, different categories are represented by different colors, and same shapes represent samples with similar semantic information.

recently [1]–[3], [26]–[28]. Most of them focus on integrating the information of multiple views into a comprehensive representation and pursuing the consistency of different views only in feature space, ignoring that the fusion of multiple views' features may cause some views with fuzzy clustering structure to interfere with the performance of the final representation, which may decline the performance of the model with the increase of the number of multi-view data views.

Aiming at these problems, some research on the non-fusion MVC method appeared [25], [26], [29], [30]. In order to obtain consistent cluster assignments of different views in feature space for clustering, most of them align the cluster assignments of different views to the cluster assignment of the common feature. Compared with fusing multiple views into a complete representation, the non-fusion model can avoid the negative impact of the view with a fuzzy cluster structure. As shown in Figure 1(a), the two separated views have clear cluster assignment mapping in the global space, so view 2 with a clear cluster structure can guide the cluster separation of view 1 by using the cluster assignments collaboration of different views. The existing non-fusion MVC methods basically focus on the alignment of cluster assignments in feature space. As shown at the top of Figure 1(b), although view 1 can be separated into three clusters through collaborative training of other views' cluster assignments, it is difficult to separate some overlapping areas in low-dimensional feature space because the computational essence of cluster assignment is mapping the distance between the low-dimensional features of views and their respective cluster centers into pseudo-labels [24],
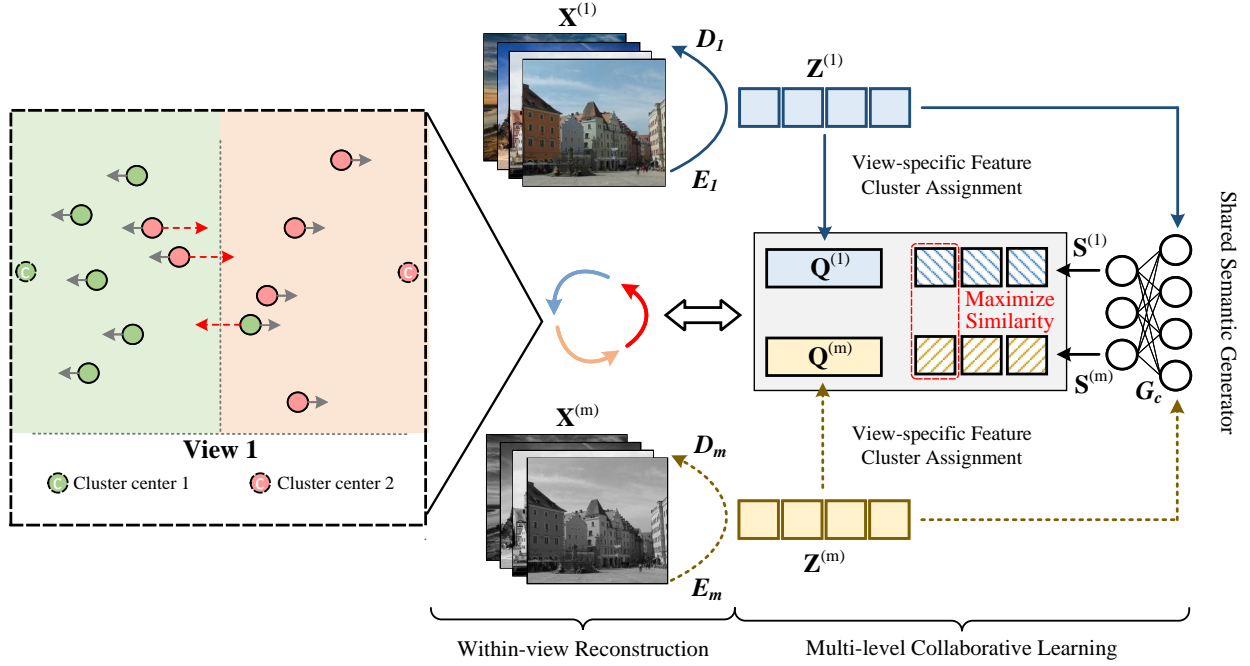
Fig. 2. Overview of MCoCo. Given m views, each view $\mathbf{X}^{(m)}$ is mapped to its feature space $\mathbf{Z}^{(m)}$ by encoder $\boldsymbol{E}_m$ and decoder $\boldsymbol{D}_m$. $\mathbf{Z}^{(m)}$ is utilized to construct the cluster assignment $\mathbf{Q}^{(m)}$ by Eq.9. Based on a shared semantic generator $\boldsymbol{G}_c$, $\mathbf{Z}^{(m)}$ is mapped into the semantic space $\mathbf{S}^{(m)}$, where the semantic labels $\{\mathbf{S}^{(i)}\}_{i=1}^m$ tend to be consistent by contrastive learning. Then, $\mathbf{Q}^{(m)}$ and $\mathbf{S}^{(m)}$ collaborate with each other to mine the multi-level consistent information. The left part illustrates the influence of one view under multi-level collaborative learning: the gray arrow indicates that the cluster assignments $\{\mathbf{Q}^{(i)}\}_{i=1}^m$ are jointly learned to make the sample closer to the cluster center, and the red arrow indicates the process of multi-level collaboration to correct the misassigned samples, where the samples with similar semantic information are forced to be close to each other.

[31], [32].

To solve the problems in the above discussion, we present a novel Multi-level Consistency Collaborative learning framework (MCoCo). As illustrated in Figure 2, MCoCo consists of two modules, namely, within-view reconstruction and multi-level collaborative learning. We consider that the data of different views describe the same object, which makes it reasonable to exploit the consensus in the semantic space. Further, in the multi-level collaborative learning module, a well-designed multi-level consistency collaboration strategy uses consistent semantic information to help the clustering assignments in feature space in a self-supervised manner. As shown in Figure 1(b), suppose that there are three categories in the dataset. Different categories are represented by different colors, and the same shape indicates that the samples with similar semantic information. In feature space, we jointly learn the cluster assignments of each view, while in semantic space, the semantic labels of different views tend to be similar through contrastive learning. Then, as shown in the middle of Figure 1(b), under the collaboration of the semantic labels the misassigned samples can be corrected. The consistent information of different levels of spaces collaborate with each other, which enables MCoCo to learn discriminating clustering assignments and mine the multi-level consistent information in multiple views.

The main contributions of the proposed method can be summarized as follows:

- We introduce a novel Multi-level Consistency Collaborative learning framework (MCoCo) for multi-view clus-

tering, which can mine multi-level consistent information to guide clustering under the collaboration of consistent information in different levels of spaces.
- We propose a brand-new multi-level consistency collaboration strategy that allows MCoCo to achieve the respective consistency goals of feature space and semantic space while realizing multi-level space collaboration.
- Extensive experiments are conducted on diverse benchmark datasets, and experimental results demonstrate its state-of-the-art clustering performance.

II. RELATED WORK

1) Deep Single-view Clustering: The emergence of deep neural networks has led to rapid development in various fields of computer science, and deep networks have been used by researchers to handle clustering tasks. The most representative work in deep single-view clustering is [31], which is an end-to-end learning method that can automatically learn feature representations and cluster assignments for data. It maps data to a low-dimensional feature space and iteratively optimizes a clustering objective to achieve simultaneous learning of feature representations and cluster assignments. Subsequently, many variant versions emerged in the continuation of this work [33], [34]. [35] is a new mechanism proposed for clustering using GANs. This is achieved by sampling from a mixture of one-hot encoded variables and continuous latent variables, and training an inverse network which projects the data to the clustering latent space. These methods have achieved impressive progress in the field of deep single-view clustering. However, they can

only handle individual views and cannot effectively leverage the rich information provided by multi-view data to enhance clustering performance.

2) Multi-view Clustering: Multi-view clustering is a challenging and important branch of multi-view learning research that aims to explore the rich semantic information in multiple views and use this information to guide clustering in low-dimensional space [36], [37]. For multi-view clustering, the key is to explore the consistency information among multiple views. CCA-based methods [38]–[41] aim to maximize the canonical correlation between different views to uncover the consistent information among them. For two views, the paradigm of CCA-based methods can be summarized as follows:

$$\min_{\beta_1, \beta_2} -corr(\boldsymbol{f}_1(\mathbf{X}^{(1)}; \beta_1), \boldsymbol{f}_2(\mathbf{X}^{(2)}; \beta_2)) + \lambda reg(\beta_1, \beta_2), \quad (1)$$

where $\boldsymbol{f}_1(\cdot; \beta_1)$ and $\boldsymbol{f}_2(\cdot; \beta_2)$ are two embedding strategies with parameters β_1 and β_2. $corr(\cdot)$ and $reg(\cdot)$ indicate the canonical correlation function and the regularization term respectively. In [41] $\boldsymbol{f}_1(\cdot; \beta_1)$ and $\boldsymbol{f}_2(\cdot; \beta_2)$ are both deep neural networks. As for [38] $\boldsymbol{f}_1(\cdot; \beta_1)$ and $\boldsymbol{f}_2(\cdot; \beta_2)$ are utilized to learn bottleneck representations of two autoencoders. In addition, some CCA-based methods [42]–[44] also consider utilizing graph information to guide clustering.

More, [45] integrates low-dimensional embedding representations and applies low-rank tensor constraints on the subspace representations of multiple views to construct a comprehensive feature representation, incorporating rich information across the views. [46] is capable of generating a unified multi-view spectral representation through the introduction of an orthogonal constraint and reformulation strategy that utilizes Cholesky decomposition during the learning process. This approach enables the model to effectively capture and exploit information across multiple views. [47] relies on the information bottleneck principle to integrate shared representation among different views and view-specific representation of each view, promoting a comprehensive representation of multi-views and flexibly balancing the complementarity and consistency among multiple views. [48] is a novel multi-view clustering method that uses m sub-cluster centers to reveal the sub-cluster structure in multi-view data, thereby improving clustering performance. To fully utilize complementary information between different views, it uses a multi-view combination weights strategy to automatically assign weights and properly fuse information from different views to obtain an optimally shared bipartite graph. [28] and [18] explored high-order correlations and graph information in multiple views in a shared representation space, respectively, and achieved equally encouraging results.

Most existing methods focus on integrating multiple views into a shared low-dimensional space for clustering, while this paper differs from the majority of methods by utilizing consistent information across different views at multiple levels to guide learning of a discriminative clustering assignment. The non-fusion strategy can avoid negative impacts from views with ambiguous clustering assignments during fusion.

3) Contrastive Learning: Contrastive learning [49], [50] is one of most effective unsupervised representation learning

TABLE I
MAIN SYMBOLS USED IN THIS PAPER.

Symbol	Meaning
m	The number of views.
N	The number of samples.
$\mathbf{X}^{(i)}$	The original feature representation in i view.
$\mathbf{Z}^{(i)}$	The view-specific feature representation in i view.
$\mathbf{S}^{(i)}$	The semantic label of the i-th view.
$\mathbf{Q}^{(i)}$	The clustering allocation distribution of the i-th view.
D_Z	The dimensionality of $\mathbf{Z}^{(i)}$
D_i	The dimensionality of $\mathbf{X}^{(i)}$

paradigm that aims to minimize the spatial distance or maximize the similarity between positive pairs, while maximizing the spatial distance between them and their corresponding negative pairs. In recent years, contrastive learning has made remarkable progress in representation learning and computer vision, such as [51] and [52]. With the utilization of deep networks in multi-view clustering, contrastive learning has also been widely used in multi-view clustering work, such as [27], [53]–[55]. [27] is base on information theory, which maximizes the mutual information between different views through contrastive learning, and solves the problem of view missing through a bidirectional prediction network. A end to end online image clustering method was proposed in [53], where contrastive learning was used to explore the consistency information between clustering space and instance space. A multi-view representation learning method [56] is proposed to apply contrastive learning for solving graph classification problems. [26] optimizes the objectives of different feature spaces separately through contrastive learning, and solves the conflict between view reconstruction loss and consistency loss.

III. METHOD

In this section, we introduce the proposed method, termed Multi-level Consistency Collaborative Multi-view Clustering (MCoCo). A multi-view dataset $\{\mathbf{X}^{(i)}\}_{i=1}^{m}$ with m view, the i-th view is denoted by $\mathbf{X}^{(i)} \in \mathbf{R}^{N \times D_i}$, where N denotes the number of samples and D_i represents the dimension of the view. In order to be more clear and concise, we have listed the main symbols used in this article in Table I. MVC aims to partition the examples into k clusters.

A. Loss Function

The loss function of MCoCo can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{Re} + \mathcal{L}_{Co}, \quad (2)$$

where $\mathcal{L}_{Re}$ is the loss of within-view reconstruction and $\mathcal{L}_{Co}$ denotes multi-level collaborative learning loss.

B. Within-view reconstruction

Generally, the dimensions and input forms of different views of multi-view data are quite different, and at the same time,

Algorithm 1 Optimization algorithm of MCoCo
***
Input: Multi-view dataset $\{\mathbf{X}^{(i)}\}_{i=1}^m$; Parameter τ; Number of categories k.

Output: Cluster assignment $\mathbf{Y}$.
1: Initialize $\{\theta_i, \phi_i\}_{i=1}^m$ by minimizing Eq.5;
2: Initialize views' cluster centroids $\{\mu^{(i)}\}_{i=1}^m$ by k-means.
3: **while** not converged **do**
4:   Obtain the view-specific representation $\{\mathbf{Z}^{(i)}\}_{i=1}^m$ through Eq.3;
5:   Obtain semantic labels $\{\mathbf{S}^{(i)}\}_{i=1}^m$;
6:   Obtain cluster assignments $\{\mathbf{Q}^{(i)}\}_{i=1}^m$ and it's target distribution $\{\mathbf{P}^{(i)}\}_{i=1}^m$ through Eq.9 and Eq.10;
7:   Obtain the target distribution $\{\mathbf{S}'^{(i)}\}_{i=1}^m$ of semantic space through Eq.11;
8:   Update $\{\theta_i, \phi_i\}_{i=1}^m$, φ and $\{\mu^{(i)}\}_{i=1}^m$ with Eq.2;
9: **end while**
***

there may be some redundant information in the original data. To learn a reliable representation for each view, we map the data of different views into a low-dimensional feature space by inputting the data $\mathbf{X}^{(i)}$ into the respective encoder $\boldsymbol{E}_i(\cdot; \theta_i)$ with parameter θ_i:

$$\mathbf{Z}_j^{(i)} = \boldsymbol{E}_i(\mathbf{X}_j^{(i)}; \theta_i), \tag{3}$$

where $\mathbf{X}_j^{(i)}$ is the j-th sample of $\mathbf{X}^{(i)}$ and $\mathbf{Z}_j^{(i)} \in \mathbf{R}^{D_Z}$ denotes the representation in the D_Z-dimensional feature space. Then we input this low-dimensional feature into the decoder $\boldsymbol{D}_i(\cdot; \phi_i)$ with parameter ϕ_i for reconstruction:

$$\hat{\mathbf{X}}_j^{(i)} = \boldsymbol{D}_i(\mathbf{Z}_j^{(i)}; \phi_i), \tag{4}$$

where $\hat{\mathbf{X}}_j^{(i)}$ is the reconstructed sample. By minimizing the flowing reconstruction loss $\mathcal{L}_{Re}$, we can transform the input $\mathbf{X}^{(i)}$ into the representation $\mathbf{Z}^{(i)}$:

$$\mathcal{L}_{Re} = \sum_{i=1}^m \sum_{j=1}^N ||\mathbf{X}_j^{(i)} - \boldsymbol{D}_i(\boldsymbol{E}_i(\mathbf{X}_j^{(i)}; \theta_i); \phi_i)||_2^2. \tag{5}$$

C. Multi-level Collaborative Learning

Based on the within-view reconstruction, we can obtain the low-dimensional representations $\{\mathbf{Z}^{(i)}\}_{i=1}^m$ of different views. In order to use the multi-level consistency information of multi-view data to guide clustering, MCoCo achieves respective consistency goals in semantic and feature space, and makes multi-level spaces collaborate with each other.

Overall, the loss function $\mathcal{L}_{Co}$ of this section consists of two parts:

$$\mathcal{L}_{Co} = \lambda_1 \mathcal{L}_{Se} + \lambda_2 \mathcal{L}_{Ml}, \tag{6}$$

where $\mathcal{L}_{Se}$ is the loss of semantic consistency, $\mathcal{L}_{Ml}$ indicates the multi-level consistency loss. Regarding λ_1 and λ_2, they are two trade-off parameters.

Contrastive learning of semantic consistency

Because the data of different views describe the same object, different views should have similar semantic labels. Then the aligned semantic labels can be used as a self-supervised signal to modify the cluster assignments in feature space, which

makes the obtained cluster assignments more discriminating. In this section, we explain how to obtain consistent semantic labels. Specifically, We obtain the semantic labels of each view $\{\mathbf{S}^{(i)} \in \mathbf{R}^{N \times k}\}_{i=1}^m$ by inputing $\{\mathbf{Z}^{(i)}\}_{i=1}^m$ into a shared semantic generator $\boldsymbol{G}_c(\cdot; \varphi)$ with the parameter φ, which is constructed by fully connected neural networks. The $\mathbf{S}_{ij}^{(m)}$ represents the probability that the i-th sample in view m belongs to the j-th class. To effectively excavate the variable semantic consensus information in semantic space and make semantic labels of different views tend to be consistent, the contrastive learning of semantic consistency is introduced here. For $\mathbf{S}_{\cdot j}^{(m)}$, there are $(mk - 1)$ column vector pairs $\{\mathbf{S}_{\cdot j}^{(m)}, \mathbf{S}_{\cdot c}^{(w)}\}_{c=1,\ldots,k}^{w=1,\ldots,m}$, of which $\{\mathbf{S}_{\cdot j}^{(m)}, \mathbf{S}_{\cdot j}^{(w)}\}_{w \neq m}$ can form $(m - 1)$ positive pairs and the remaining $m(k - 1)$ pairs form negative pairs. The cosine similarity is utilized to measure the similarity between two semantic column vectors:

$$d(\mathbf{S}_{\cdot j}^{(i)}, \mathbf{S}_{\cdot c}^{(j)}) = \frac{\mathbf{S}_{\cdot i}^{(i)} \cdot \mathbf{S}_{\cdot c}^{(j)}}{||\mathbf{S}_{\cdot i}^{(i)}|| ||\mathbf{S}_{\cdot c}^{(j)}||}. \tag{7}$$

We define the semantic consistency loss $l(i, j)$ between $\mathbf{S}^{(i)}$ and $\mathbf{S}^{(j)}$ as:

$$-\frac{1}{k} \sum_{c=1}^k \log \frac{e^{d(\mathbf{S}_{\cdot c}^{(i)}, \mathbf{S}_{\cdot c}^{(j)})/\tau}}{(\sum_{w=1}^k (e^{d(\mathbf{S}_{\cdot c}^{(i)}, \mathbf{S}_{\cdot w}^{(i)})/\tau} + e^{d(\mathbf{S}_{\cdot c}^{(i)}, \mathbf{S}_{\cdot w}^{(j)})/\tau}) - e^{\frac{1}{\tau}}},$$

where τ is the temperature parameter. The complete contrastive learning loss of semantic consistency can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{Se} = &\frac{1}{2} \sum_{i=1}^m \sum_{j=1, j \neq i}^m l(i, j) \\ &+ \sum_{i=1}^m \sum_{c=1}^k (\frac{1}{N} \sum_{j=1}^N \mathbf{S}_{jc}^{(i)} \log \frac{1}{N} \sum_{j=1}^N \mathbf{S}_{jc}^{(i)}).\end{aligned} \tag{8}$$

The second part of Eq.8 is a regularization term, which can avoid grouping all samples into the same cluster.

Multi-level consistency collaboration

In order to obtain the clustering assignments of each view in feature space, we initialize the learnable parameters $\{\mu_j^{(i)} \in \mathbf{R}^{D_Z}\}_{j=1}^k$ by k-means [31], where $\mu_j^{(i)}$ represents the j-th cluster centroid of the i-th view. According to [24], [29], [31], we use Student's t-distribution to generate soft cluster assignments, which can be described as:

$$\mathbf{Q}_{ij}^{(m)} = \frac{(1 + ||\mathbf{Z}_i^{(m)} - \mu_j^{(m)}||^2)^{-1}}{\sum_j (1 + ||\mathbf{Z}_i^{(m)} - \mu_j^{(m)}||^2)^{-1}}, \tag{9}$$

where $\mathbf{Q}_{ij}^{(m)}$ is treated as the pseudo label that represents the probability of assigning the i-th sample of the m-th view to the j-th category. The higher probability of pseudo label means that the high probability components of pseudo label has higher confidence. In order to increase the discrimination ability of pseudo label with higher confidence, we enhance $\mathbf{Q}_{ij}^{(m)}$ to an auxiliary target distribution $\mathbf{P}^{(m)}$ with the operation of square and normalization:

$$\mathbf{P}_{ij}^{(m)} = \frac{(\mathbf{Q}_{ij}^{(m)})^2 / \sum_i \mathbf{Q}_{ij}^{(m)}}{\sum_j ((\mathbf{Q}_{ij}^{(m)})^2 / \sum_i \mathbf{Q}_{ij}^{(m)})}. \tag{10}$$

TABLE II
THE INFORMATION OF THE DATASETS IN OUR EXPERIMENTS.

Dataset	#Sample	#Cluster	#View	#Dimensionality of features
MNIST-USPS	5000	10	2	{784, 784}
Multi-COIL-20	1440	20	3	{1024, 1024, 1024}
BDGP	2500	5	2	{1750, 79}
Multi-MNIST	70000	10	2	{1024, 1024}
Multi-Fashion	10000	10	3	{784, 784, 784}
Noisy-MNIST	50000	10	2	{1024, 1024}
Caltech-2V	1400	7	2	{40, 254}
Caltech-3V	1400	7	3	{40, 254, 928}
Caltech-4V	1400	7	4	{40, 254, 928, 512}
Caltech-5V	1400	7	5	{40, 254, 928, 512, 1984}

In order to make multi-views collaborate with each other to get consistent cluster assignments and, at the same time, make use of the aligned semantic labels to weakly supervise the cluster assignments in feature space, we propose a brand-new multi-level consistency collaboration strategy, which can achieve multi-level collaboration. Specifically, we enhance the semantic labels of each view according to Eq.10 to get the target distribution of semantic space:

$$\mathbf{S}_{ij}^{'(m)} = \frac{(\mathbf{S}_{ij}^{(m)})^2/\sum_i \mathbf{S}_{ij}^{(m)}}{\sum_j((\mathbf{S}_{ij}^{(m)})^2/\sum_i \mathbf{S}_{ij}^{(m)})}. \quad (11)$$

As thus, the multi-level consistency loss $\mathcal{L}_{Ml}$ is defined as:

$$\mathcal{L}_{Ml} = \sum_{k=1}^{m}(\sum_{c=1}^{m}(D_{kl}(\mathbf{P}^{(c)}||\mathbf{Q}^{(k)})) + D_{kl}(\mathbf{S}^{'(k)}||\mathbf{Q}^{(k)}))$$
$$= \sum_{k=1}^{m}\sum_{i=1}^{N}\sum_{j=1}^{k}(\sum_{c=1}^{m}(\mathbf{P}_{ij}^{(c)} \log \frac{\mathbf{P}_{ij}^{(c)}}{\mathbf{Q}_{ij}^{(k)}}) + \mathbf{S}_{ij}^{'(k)} \log \frac{\mathbf{S}_{ij}^{'(k)}}{\mathbf{Q}_{ij}^{(k)}}), \quad (12)$$

where D_{kl} indicates the Kullback-Leibler divergence. By optimizing Eq.12, different views jointly learned with each other in the feature space, and the samples with similar semantic information attract each other. In this way, MCoCo can mine multi-level consistency information to learn discriminating clustering assignments.

MCoCo's multi-level consistency collaboration strategy enables it to obtain the consistent cluster assignments of multiple views. In order to avoid the interference of a few false predictions and realize clear cluster assignments with high confidence, the final cluster assignment is calculated as follows:

$$\mathbf{Y}_i = \arg \max_j(\frac{1}{m}\sum_{k=1}^{m}\mathbf{Q}_{ij}^{(k)}). \quad (13)$$

For clarification, the optimization procedure of MCoCo is summarized in Algorithm 1.

IV. EXPERIMENTS

To verify the effectiveness of our method, extensive experiments are conducted in this section. Furthermore, detailed discussions of our method are provided as well.

A. Experiments Setup

Datasets. The benchmark datasets we used in our experiments are shown in Table II:

1) MNIST-USPS [17]: It is a two-view dataset that contains 5000 handwritten digital image samples from numbers 0 to 9.

2) Multi-COIL-20 [57]: It is a three-view dataset containing 1440 pictures of 20 categories, and different views represent different poses of the same object.

3) BDGP [58]: It is a two-view dataset containing 2500 images of drosophila embryos belonging to 5 categories, each with visual and textual features.

4) Multi-MNIST [25]: It contains 70000 handwritten digital images belonging to 10 classes, which have two views, and different views imply the same digit written by different people.

5) Multi-Fashion [59]: It has 10000 images collected from 10 categories about fashion products and has three views. We use the same data set as [29], which randomly selects a sample with the same label from this set to construct the second and third view.

6) Noisy-MNIST [27]: It uses the original 70000 MNIST images as the first view and randomly selects within-class images with white Gaussian noise as the second view. We followed the setting of article [22] and used a subset of Noisy-MNIST containing 50000 samples.

7) Caltech-nV [60]: It is an RGB image dataset with multiple views, which contains 1400 images belonging to 7 categories and have five views. Four sub-datasets, namely Caltech-2V, Caltech-3V, Caltech-4V, and Caltech-5V, with different numbers of views, are built for evaluating the robustness of the comparison methods in terms of the number of views. Specifically, Caltech-2V uses WM and CENTRISTT; Caltech-3V uses WM, CENTRIST, and LBP; Caltech-4V uses WM, CENTRIST, LBP, and GIST; Caltech-5V uses WM, CENTRIST, LBP, GIST, and HOG.

Evaluation metrics. Four metrics are utilized to evaluate the clustering quality, i.e., Accuracy (ACC), Normalized Mutual Information (NMI), Rand Index (RI), and Fscore. To eliminate the randomness and make the experimental results more reliable, we take ten trials for all experiments.

TABLE III
CLUSTERING RESULTS ON SMALL-SCALE DATASETS. THE BEST AND SECOND BEST RESULTS HAVE BEEN MARKED IN BOLD AND UNDERLINED RESPECTIVELY.

Datasets	MNIST-USPS				Multi-COIL-20				BDGP			
Metrics	ACC	NMI	Fscore	RI	ACC	NMI	Fscore	RI	ACC	NMI	Fscore	RI
LMSC(2017)	0.373	0.433	0.404	0.771	0.683	0.766	0.613	0.954	0.524	0.432	0.556	0.604
AE^2-Nets(2019)	0.626	0.623	0.567	0.903	0.740	0.862	0.771	0.964	0.552	0.406	0.501	0.661
DEMVC(2021)	0.901	0.930	0.897	0.979	<u>0.825</u>	<u>0.935</u>	<u>0.871</u>	0.980	0.609	0.529	0.557	0.818
DUA-Nets(2021)	0.751	0.689	0.930	0.660	0.602	0.711	0.597	0.946	0.603	0.406	0.539	0.762
DCP(2022)	0.891	0.941	<u>0.928</u>	0.976	0.690	0.887	0.621	0.958	0.438	0.385	0.534	0.542
SDMVC(2022)	<u>0.937</u>	<u>0.943</u>	0.913	<u>0.983</u>	0.809	0.905	0.823	<u>0.981</u>	<u>0.965</u>	<u>0.909</u>	<u>0.934</u>	<u>0.961</u>
CMRL(2023)	0.916	0.856	0.860	0.971	0.792	0.878	0.797	0.977	0.789	0.672	0.826	0.717
MCoCo(ours)	**0.995**	**0.986**	**0.998**	**0.990**	**0.999**	**0.999**	**0.999**	**0.999**	**0.987**	**0.959**	**0.989**	**0.972**

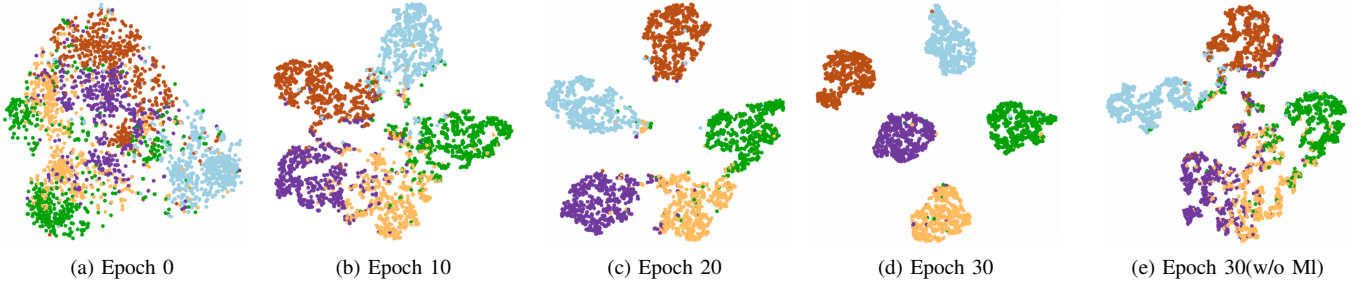| (a) Epoch 0 | (b) Epoch 10 | (c) Epoch 20 | (d) Epoch 30 | (e) Epoch 30(w/o Ml) |

Fig. 3. The t-SNE visualization results of the BDGP dataset (view one) in Epoch 0, 10, 20, and 30. (e) showing the representation of view one learned by MCoCo at epoch 30 without multi-level collaboration.

Comparison methods. The following state-of-the-art algorithms are used for comparison:

1) LMSC [61]: It learns the latent unified representation by mapping different views' view-specific feature into a common space and employing the low-rank subspace constraint.

2) AE^2-Nets [3]: Nested autoencoders are used to learn the compact unified representation by balancing the complementarity and consistency among multiple views.

3) DEMVC [22]: It proposed a non-fusion model of collaborative training among cluster assignments of multiple views.

4) DUA-Nets [1]: It presents the dynamic uncertainty-aware networks for UMRL. By estimating and leveraging the uncertainty of data, it achieves the noise-free multi-view feature representation.

5) DCP [62]: It learns the unified multi-view representation by maximizing the mutual information of different views via contrastive learning in the feature space.

6) CMRL [45]: It introduces the orthogonal mapping strategy and imposing the low-rank tensor constraint on the subspace representations.

7) SDMVC [29]: It concatenates the features of multiple views as global feature and uses global discriminative information to supervise all views to learn more discriminative view-specific features.

Implementation details. For all datasets, the ReLU [63] activation function is used to implement autoencoders in MCoCo. Adam optimizer [64] is employed for optimization. Our method is implemented by PyTorch [65] on one NVIDIA Geforce GTX 2080ti GPU with 11GB memory.

B. Experimental Result

We discuss the clustering performance of MCoCo compared with other state-of-art algorithms on three different datasets: small-scale datasets, large-scale datasets, and datasets with a variable number of views. Generally speaking, the proposed method can achieve the best performance in all cases.

The results on small-scale and large-scale datasets are reported in Tables III and IV. We can observe that MCoCo has achieved the best performance on all metrics, whether it is a large-scale dataset or a small-scale dataset. Compared with the second-best method, MCoCo's ACC, NMI and Fscore are all improved by more than 10% on Multi-COIL-20, Multi-Fashion, and Noisy-MNIST. The main reason is that MCoCo makes different levels of spaces collaborate with each other while achieving their own consistency goals. In this way, MCoCo can fully mine the multi-level consistent information of different views, which is utilized to guide the process of clustering. For Noisy-MNIST, the second view has a chaotic clustering structure because of the white Gaussian noise. If multiple views are fused or mapped to the same space, this may cause the private information in the second view to have a negative impact on the final clustering effect. Compared with AE^2-Nets, DCP, and DUA-Nets, which integrate multiple views into a unified representation for clustering, MCoCo can avoid the negative impact of private information of views with white Gaussian noise. At the same time, it can be seen that compared with traditional clustering methods based on matrix decomposition, such as CMRL and LMSC, MCoCo has the advantage of complexity in dealing with the clustering of

Datasets	Multi-MNIST				Multi-Fashion				Noisy-MNIST			
Metrics	ACC	NMI	Fscore	RI	ACC	NMI	Fscore	RI	ACC	NMI	Fscore	RI
LMSC(2017)	-	-	-	-	0.439	0.417	0.367	0.828	-	-	-	-
AE^2-Nets(2019)	0.737	0.645	0.624	0.924	0.729	0.763	0.716	0.935	0.220	0.098	0.162	0.813
DEMVC(2021)	0.996	0.997	0.996	0.996	0.720	0.848	0.778	0.926	0.589	0.714	0.633	0.900
DUA-Nets(2021)	0.795	0.742	0.713	0.943	0.772	0.761	0.727	0.945	0.179	0.066	0.145	0.808
DCP(2022)	0.793	0.905	0.747	0.952	0.757	0.862	0.822	0.948	0.786	0.883	0.740	0.921
SDMVC(2022)	0.998	0.996	0.998	0.998	0.860	0.876	0.845	0.965	0.557	0.527	0.460	0.887
CMRL(2023)	-	-	-	-	0.768	0.803	0.748	0.943	-	-	-	-
MCoCo(ours)	**0.999**	**0.998**	**0.999**	**0.999**	**0.991**	**0.977**	**0.996**	**0.982**	**0.994**	**0.981**	**0.998**	**0.988**

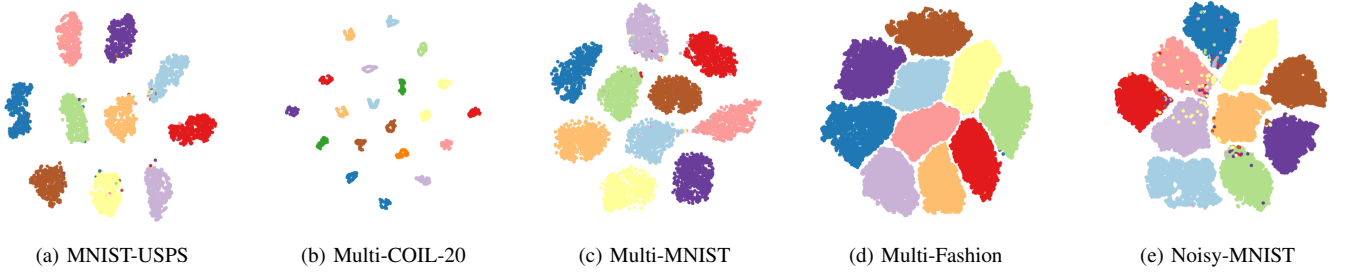| (a) MNIST-USPS | (b) Multi-COIL-20 | (c) Multi-MNIST | (d) Multi-Fashion | (e) Noisy-MNIST |

Fig. 4. The t-SNE visualization results of view one in MNIST-USPS, Multi-COIL-2O, Multi-MNIST, Multi-Fashion and Noisy-MNIST.
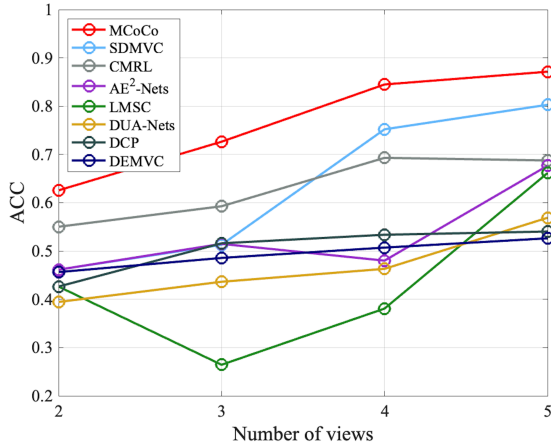

Fig. 5. Clustering results on Caltech-nV.

large-scale datasets.

In order to further verify our method, we conducted experiments on datasets with a variable number of views. Figure 5 shows the clustering results on Caltech-nV with different views. We can see that the clustering performance of MCoCo grows steadily when the number of views increases. At the same time, compared with AE^2-Nets, CMRL, and LMSC, which are methods that need to fuse different views, MCoCo can effectively avoid the negative impact of views with chaotic clustering structure on clustering results. All these indicate that MCoCo can effectively mine different levels of consistent information, and the way of non-fusion can reduce the negative impact of unclear views on clustering.

C. Visualization Result

To vividly reveal the structure of the low-dimensional representation $\mathbf{Z}^{(m)}$, we visualize it achieved in Epoch 0, 10, 20, and 30 of the MCoCo learning process based on the t-SNE. The visualization results are shown in Figure 3. From Figure 3(d), we can see that the $\mathbf{Z}^{(m)}$ with a promising structure can be achieved by our method. As can be seen from Figure 3(e), if the multi-level collaboration strategy is canceled, the clusters in the overlapping area can't be separated well.

To better demonstrate MCoCo's ability to separate clusters and obtain discriminative representations for each view on multiple datasets, we present in Figure 4 the representations obtained by MCoCo for the first view of the remaining datasets in Table III and Table IV.

V. MODEL ANALYSIS

A. Ablation Studies

In order to verify the effectiveness of each part of our method, we conduct ablation studies here. Consequently, we discuss the learning process of our method with and without $\mathcal{L}_{Se}$ and $\mathcal{L}_{Ml}$. Especially for canceling $\mathcal{L}_{Ml}$, we mean canceling the second half of Eq.12, in other words, canceling multi-level collaboration. We take the experiments on the BDGP dataset as an example. The clustering results in metrics of ACC and NMI are reported in Table V. From the experimental results, we can find that: (1) In our proposed method, both $\mathcal{L}_{Se}$ and $\mathcal{L}_{Ml}$ can effectively improve the clustering performance; (2) Compared with the original version, the ACC of the complete MCoCo can be improved by 30%, which shows that multi-level collaboration is very important for clustering

(a) ACC in clustering task   (b) NMI in clustering task   (c) Sensitivity of τ
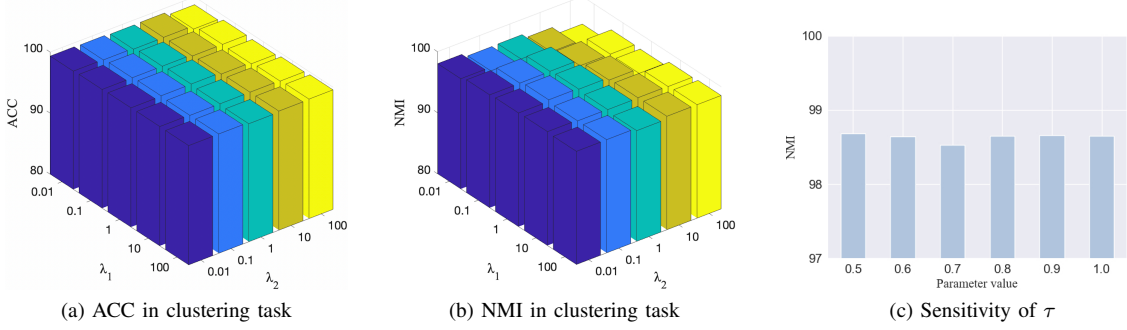
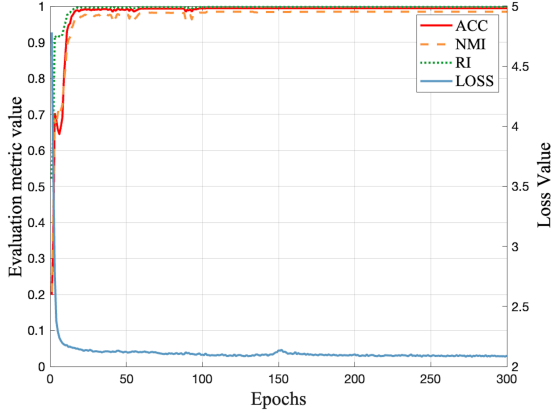Fig. 6. Parameter sensitivity analysis of MCoCo on MNIST-USPS dataset.

Fig. 7. Convergence curve and clustering performance on MNIST-USPS dataset. The X axis denotes training epochs, and the left and right Y axis denote the evaluation metric value and corresponding loss value, respectively.

TABLE V
ABLATION STUDY ON BDGP DATASET. IN WHICH (VIEW 1) OR (VIEW 2) INDICATES THAT THE FINAL RESULT IS ONLY OBTAINED BY CLUSTER ASSIGNMENTS OF VIEW1 OR CLUSTER ASSIGNMENTS OF VIEW2.

Method	Components		Metrics	
	$\mathcal{L}_{Se}$	$\mathcal{L}_{Ml}$	ACC	NMI
MCoCo(view 1)			0.9848	0.9484
MCoCo(view 2)	✓	✓	0.9800	0.9479
MCoCo			**0.9872**	**0.9592**
MCoCo(view 1)			0.8120	0.7821
MCoCo(view 2)	✓		0.8092	0.7839
MCoCo			**0.8152**	**0.7984**
MCoCo(view 1)			0.6716	0.6488
MCoCo(view 2)		✓	0.6696	0.6478
MCoCo			**0.6840**	**0.6543**

tasks; (3) According to the clustering performance of MCoCo (view1) and MCoCo (view2), we can find that MCoCo can align different views well.

More specifically, it can be seen from Figure 3: According to Figure 3(a), it can be known that the different clusters in the first view of the BDGP dataset overlap seriously, and there are only two views in BDGP dataset. It is difficult to collaboratively separate these overlapping clusters through another view. If there is no collaboration of semantic labels in semantic space, the result will be as shown in Figure 3(e). Different clusters in Figure 3(e) are difficult to separate, resulting in poor clustering performance.

B. Parameter Sensitivity Analysis

To explore the sensitivity of MCoCo to hyper-parameters, we first conducted an experiment on the MNIST-USPS dataset. In the experiment, we set different values to λ_1 and λ_2 in Eq.6 and explore their influence on the clustering task in the metric of ACC and NMI. In order to eliminate the randomness of the experiment and make the experimental results more reliable, our final results are all averaged by ten times clustering experiments. The final experimental results are shown in Figure 6(a) and Figure 6(b). From the results, we can see that MCoCo is insensitive to the hyper-parameters λ_1 and λ_2. In order to ensure the uniformity of all experiments, λ_1 and λ_2 are all fixed at 1 in other experiments in this paper.

For another hyper-parameter τ in Eq.8, we set τ to $\{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ respectively, and do the same ten times for each clustering experiment to finally get the average value of metric NMI. Figure 6(c) shows the experimental results on the dataset MNIST-USPS. The results show that MCoCo is also robust to τ. Actually, for all other experiments, we fixed τ at 0.5.

C. Convergence Analysis

To show the convergence properties of MCoCo, we take the experiment on the MNIST-USPS dataset and display the experimental results in Figure 6. It can be observed that the loss value drops rapidly in the first 15 epochs, with ACC, NMI, and RI continuously increasing. For other datasets, similar convergence properties can be achieved as well.

VI. CONCLUSION

In this paper, we propose a novel Multi-level Consistency Collaborative learning framework (MCoCo) for multi-view clustering, which can fully mine multi-level consistent information to guide the process of clustering. While achieving consistency goals in different spaces, MCoCo can realize the collaboration between cluster assignments and consistent semantic labels. Therefore, our method can get more discriminating clustering assignments for clustering. Meanwhile,

MCoCo is also robust to some views with unclear clustering structure in a non-fusion manner. Experimental results on several benchmark datasets verify the effectiveness of MCoCo over other state-of-the-art methods.

VII. Acknowledgment

References

[1] Y. Geng, Z. Han, C. Zhang, and Q. Hu, "Uncertainty-aware multi-view representation learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 9, 2021, pp. 7545–7553.

[2] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, and D. Xu, "Generalized latent multi-view subspace clustering," IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 1, pp. 86–99, 2018.

[3] C. Zhang, Y. Liu, and H. Fu, "Ae2-nets: Autoencoder in autoencoder networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2577–2585.

[4] Q. Zheng, J. Zhu, Z. Li, S. Pang, J. Wang, and Y. Li, "Feature concatenation multi-view subspace clustering," Neurocomputing, vol. 379, pp. 89–102, 2020.

[5] Y. Jia, H. Liu, J. Hou, S. Kwong, and Q. Zhang, "Multi-view spectral clustering tailored tensor low-rank representation," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 12, pp. 4784–4797, 2021.

[6] M. Yin, J. Gao, S. Xie, and Y. Guo, "Multiview subspace clustering via tensorial t-product representation," IEEE transactions on neural networks and learning systems, vol. 30, no. 3, pp. 851–864, 2018.

[7] Y. Xie, J. Liu, Y. Qu, D. Tao, W. Zhang, L. Dai, and L. Ma, "Robust kernelized multiview self-representation for subspace clustering," IEEE transactions on neural networks and learning systems, vol. 32, no. 2, pp. 868–881, 2020.

[8] J. Guo, Y. Sun, J. Gao, Y. Hu, and B. Yin, "Rank consistency induced multiview subspace clustering via low-rank matrix factorization," IEEE Transactions on Neural Networks and Learning Systems, 2021.

[9] Z. Kang, W. Zhou, Z. Zhao, J. Shao, M. Han, and Z. Xu, "Large-scale multi-view subspace clustering in linear time," in Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 04, 2020, pp. 4412–4419.

[10] Y. Chen, X. Xiao, C. Peng, G. Lu, and Y. Zhou, "Low-rank tensor graph learning for multi-view subspace clustering," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 1, pp. 92–104, 2021.

[11] M. Lan, M. Meng, J. Yu, and J. Wu, "Generalized multi-view collaborative subspace clustering," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 6, pp. 3561–3574, 2021.

[12] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in Proceedings of the 2013 SIAM international conference on data mining. SIAM, 2013, pp. 252–260.

[13] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," IEEE transactions on neural networks and learning systems, vol. 29, no. 10, pp. 4833–4843, 2018.

[14] Z. Yang, N. Liang, W. Yan, Z. Li, and S. Xie, "Uniform distribution non-negative matrix factorization for multiview clustering," IEEE transactions on cybernetics, vol. 51, no. 6, pp. 3249–3262, 2020.

[15] F. Nie, J. Li, X. Li et al., "Self-weighted multiview clustering with multiple graphs." in IJCAI, 2017, pp. 2564–2570.

[16] H. Wang, Y. Yang, and B. Liu, "Gmc: Graph-based multi-view clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 6, pp. 1116–1129, 2019.

[17] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "Comic: Multi-view clustering without parameter selection," in International conference on machine learning. PMLR, 2019, pp. 5092–5101.

[18] Q. Zheng, J. Zhu, Z. Li, and H. Tang, "Graph-guided unsupervised multiview representation learning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 1, pp. 146–159, 2022.

[19] W. K. Wong, N. Han, X. Fang, S. Zhan, and J. Wen, "Clustering structure-induced robust multi-view graph recovery," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 10, pp. 3584–3597, 2019.

[20] H. Wang, G. Jiang, J. Peng, R. Deng, and X. Fu, "Towards adaptive consensus graph: Multi-view clustering via graph collaboration," IEEE Transactions on Multimedia, 2022.

[21] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," Communications of the ACM, vol. 65, no. 1, pp. 99–106, 2021.

[22] H. Xu, P. Liang, W. Yu, J. Jiang, and J. Ma, "Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators." in IJCAI, 2019, pp. 3954–3960.

[23] D. Tao, Y. Guo, B. Yu, J. Pang, and Z. Yu, "Deep multi-view feature learning for person re-identification," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2657–2666, 2017.

[24] Z. Li, Q. Wang, Z. Tao, Q. Gao, Z. Yang et al., "Deep adversarial multi-view clustering network." in IJCAI, 2019, pp. 2952–2958.

[25] J. Xu, Y. Ren, H. Tang, X. Pu, X. Zhu, M. Zeng, and L. He, "Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9234–9243.

[26] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, and L. He, "Multi-level feature learning for contrastive multi-view clustering," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16 051–16 060.

[27] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, and X. Peng, "Completer: Incomplete multi-view clustering via contrastive prediction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11 174–11 183.

[28] Q. Zheng, J. Zhu, and Z. Li, "Collaborative unsupervised multi-view representation learning," IEEE Transactions on Circuits and Systems for Video Technology, 2021.

[29] J. Xu, Y. Ren, H. Tang, Z. Yang, L. Pan, Y. Yang, X. Pu, S. Y. Philip, and L. He, "Self-supervised discriminative feature learning for deep multi-view clustering," IEEE Transactions on Knowledge and Data Engineering, 2022.

[30] J. Xu, Y. Ren, G. Li, L. Pan, C. Zhu, and Z. Xu, "Deep embedded multi-view clustering with collaborative training," Information Sciences, vol. 573, pp. 279–290, 2021.

[31] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in International conference on machine learning. PMLR, 2016, pp. 478–487.

[32] J. Cheng, Q. Wang, Z. Tao, D. Xie, and Q. Gao, "Multi-view attribute graph convolution networks for clustering," in Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 2973–2979.

[33] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation." in Ijcai, 2017, pp. 1753–1759.

[34] K. Ghasedi Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5736–5745.

[35] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan, "Clustergan: Latent space clustering in generative adversarial networks," in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 4610–4617.

[36] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," Information Fusion, vol. 38, pp. 43–54, 2017.

[37] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," arXiv preprint arXiv:1304.5634, 2013.

[38] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in International conference on machine learning. PMLR, 2015, pp. 1083–1092.

[39] H. Hotelling, "Relations between two sets of variates," Breakthroughs in statistics: methodology and distribution, pp. 162–190, 1992.

[40] S. Akaho, "A kernel method for canonical correlation analysis," arXiv preprint cs/0609071, 2006.

[41] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in International conference on machine learning. PMLR, 2013, pp. 1247–1255.

[42] J. Shao, L. Wang, Z. Zhao, A. Cai et al., "Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval," Neurocomputing, vol. 214, pp. 618–628, 2016.

[43] X. Shen, Q. Sun, and Y. Yuan, "A unified multiset canonical correlation analysis framework based on graph embedding for multiple feature extraction," Neurocomputing, vol. 148, pp. 397–408, 2015.

[44] J. Chen, G. Wang, Y. Shen, and G. B. Giannakis, "Canonical correlation analysis of datasets with a common source graph," IEEE Transactions on Signal Processing, vol. 66, no. 16, pp. 4398–4408, 2018.

[45] Q. Zheng, J. Zhu, Z. Li, Z. Tian, and C. Li, "Comprehensive multi-view representation learning," Information Fusion, vol. 89, pp. 198–209, 2023.

[46] Z. Huang, J. T. Zhou, H. Zhu, C. Zhang, J. Lv, and X. Peng, "Deep spectral representation learning from multi-view data," IEEE Transactions on Image Processing, vol. 30, pp. 5352–5362, 2021.

[47] Z. Wan, C. Zhang, P. Zhu, and Q. Hu, "Multi-view information-bottleneck representation learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 11, 2021, pp. 10 085–10 092.

[48] Y. Hu, Z. Song, B. Wang, J. Gao, Y. Sun, and B. Yin, "Akm 3 c: Adaptive k-multiple-means for multi-view clustering," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 11, pp. 4214–4226, 2021.

[49] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[50] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, Z. Li, and P. Luo, "Detco: Unsupervised contrastive learning for object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8392–8401.

[51] C. Niu, H. Shan, and G. Wang, "Spice: Semantic pseudo-labeling for image clustering," IEEE Transactions on Image Processing, vol. 31, pp. 7264–7278, 2022.

[52] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, "Scan: Learning to classify images without labels," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X. Springer, 2020, pp. 268–285.

[53] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 10, 2021, pp. 8547–8555.

[54] S. Roy and A. Etemad, "Self-supervised contrastive learning of multi-view facial expressions," in Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 253–257.

[55] F. Lin, B. Bai, K. Bai, Y. Ren, P. Zhao, and Z. Xu, "Contrastive multi-view hyperbolic hierarchical clustering," arXiv preprint arXiv:2205.02618, 2022.

[56] K. Hassani and A. H. Khasahmadi, "Contrastive multi-view representation learning on graphs," in International conference on machine learning. PMLR, 2020, pp. 4116–4126.

[57] Z. Wan, C. Zhang, P. Zhu, and Q. Hu, "Multi-view information-bottleneck representation learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 11, 2021, pp. 10 085–10 092.

[58] X. Cai, H. Wang, H. Huang, and C. Ding, "Joint stage recognition and anatomical annotation of drosophila gene expression patterns," Bioinformatics, vol. 28, no. 12, pp. i16–i24, 2012.

[59] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017.

[60] L. Fei-Fei, R. Fergus, and P. 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. IEEE, 2004, pp. 178–178.

[61] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao, "Latent multi-view subspace clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4279–4287.

[62] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng, "Dual contrastive prediction for incomplete multi-view representation learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[63] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[65] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., "Pytorch: An imperative style, high-performance deep learning library," Advances in neural information processing systems, vol. 32, 2019.