RESEARCH-ARTICLE

Subspace-Contrastive Multi-View Clustering

LELE FU, Sun Yat-Sen University, Guangzhou, Guangdong, China

SHENG HUANG, Sun Yat-Sen University, Guangzhou, Guangdong, China

LEI ZHANG, Sun Yat-Sen University, Guangzhou, Guangdong, China

JINGHUA YANG, Southwest Jiaotong University, Chengdu, Sichuan, China

ZIBIN ZHENG, Sun Yat-Sen University, Guangzhou, Guangdong, China

CHUANFU ZHANG, Sun Yat-Sen University, Guangzhou, Guangdong, China

View all

Published: 24 October 2024
Online AM: 28 June 2024
Accepted: 17 June 2024
Revised: 01 April 2024
Received: 06 March 2023

Subspace-Contrastive Multi-View Clustering

LELE FU, SHENG HUANG, and LEI ZHANG, School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou, China
JINGHUA YANG, School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China
ZIBIN ZHENG, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
CHUANFU ZHANG, School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou, China
CHUAN CHEN, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

Most multi-view clustering methods based on shallow models are limited in sound nonlinear information perception capability, or fail to effectively exploit complementary information hidden in different views. To tackle these issues, we propose a novel Subspace-Contrastive Multi-View Clustering (SCMC) approach. Specifically, SCMC utilizes a set of view-specific auto-encoders to map the original multi-view data into compact features capturing its nonlinear structures. Considering the large semantic gap of data from different modalities, we project multiple heterogeneous features into a joint semantic space, namely the embedded compact features are passed through the self-expression layers to learn the subspace representations, respectively. In order to enhance the discriminability and efficiently excavate the complementarity of various subspace representations, we use the contrastive strategy to maximize the similarity between positive pairs while differentiate negative pairs. Thus, the graph regularization is employed to encode the local geometric structure within varying subspaces for optimizing the consistent affinity matrix. Furthermore, to endow the proposed SCMC with the ability of handling the multi-view out-of-samples, we develop a consistent sparse representation (CSR) learning mechanism over the in-samples. To demonstrate the effectiveness of the proposed model, we conduct a large number of comparative experiments on ten challenging datasets, and the experimental results show that SCMC outperforms existing shallow and deep multi-view clustering methods. In addition, the experimental results on out-of-samples illustrate the effectiveness of the proposed CSR.

CCS Concepts: • **Computing methodologies** → **Cluster analysis**;

The research is supported by the National Key Research and Development Program of China (2023YFB2703700), the National Natural Science Foundation of China (62176269), the Guangzhou Science and Technology Program (2023A04J0314), the Natural Science Foundation of Sichuan Province under Grants 2024NSFSC1467, the Postdoctoral Fellowship Program of CPSF under Grant GZC20232198, the Fundamental Research Funds for the Central Universities under Grant 2682024CX017.
Authors' Contact Information: Lele Fu, School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou, China; e-mail: fulle@mail2.sysu.edu.cn; Sheng Huang, School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou, China; e-mail: huangsh253@mail2.sysu.edu.cn; Lei Zhang, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China; e-mail: zhanglei73@mail2.sysu.edu.cn; Jinghua Yang, School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China; e-mail: yangjinghua110@126.com; Zibin Zheng, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China; e-mail: zhzibin@mail.sysu.edu.cn; Chuanfu Zhang (corresponding author), School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou, China; e-mail: zhangchf9@mail.sysu.edu.cn; Chuan Chen (corresponding author), School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China; e-mail: chenchuan@mail.sysu.edu.cn.

ACM 1556-472X/2024/10-ART211
https://doi.org/10.1145/3674839

Additional Key Words and Phrases: Multi-view clustering, subspace clustering, multi-view fusion, contrastive learning

ACM Reference format:
Lele Fu, Sheng Huang, Lei Zhang, Jinghua Yang, Zibin Zheng, Chuanfu Zhang, and Chuan Chen. 2024. Subspace-Contrastive Multi-View Clustering. *ACM Trans. Knowl. Discov. Data.* 18, 9, Article 211 (October 2024), 35 pages.
https://doi.org/10.1145/3674839

1 Introduction

With the growing popularity of data generation and feature extraction, multi-view or multimedia data are available in large quantities. To be specific, multi-view data refer to various feature representations from multiple aspects of objects. For instance, an image can be characterized by **Wavelet Texture (WT)**, **Local Binary Pattern (LBP)**, **Histogram of Oriented Gradient (HOG)**, and so forth. A piece of document can be expressed in numerous languages. Researchers generally believe that multi-view data consist of rich and useful heterogeneous information, so the technologies related to multi-view analysis [8, 35, 54, 62, 68] are receiving increasing attention. **Multi-View Clustering (MVC)** [15] is one of the representative technologies, which aims to explore the complementary and consistent information embedded in multi-view data to boost the clustering performance.

Currently, there are extensive MVC methods. For example, graph-based MVC [40, 46, 70] learned the connectivity graph matrices to reveal the relationship of samples, then the designed fusion schemes were developed to merge these graph matrices into a global graph. Spectral embedding-based MVC [14, 28, 47] exploited low-dimensional spectral embedding with orthogonal constraint for each view, which portrayed important components of data, then a consensus representation was further merged. The goal of nonnegative matrix based MVC [21, 22, 61] was to factorize a nonnegative discrete cluster indicator matrix from varying representations, thus the $argmax(\cdot)$ function was adopted to acquire the data labels without post-processing. Among multitudinous MVC methods, multi-view subspace clustering is a research hotspot and widely studied for its superior performance, which absorbs theory from conventional subspace clustering [3] and further develops it. The works [16, 30, 36] were classic multi-view subspace clustering approaches, which aimed to explore a uniform underlying subspace representation from multiple feature spaces. These shallow models have yielded promising clustering results, but most real-world data are high-dimensional and nonlinear, shallow models might not been equipped with the ability to fetch nonlinear structures.

Auto-Encoder (AE) is an effective unsupervised deep representation learning paradigm, which non-linearly maps the original data features into a compact feature space via the encoders, then passes the compact representations through the decoders to reconstruct the data. AE is frequently used to condense data information in clustering tasks. [19, 56] were two well-known deep embedding learning methods, which used Kullback–Leibler divergence regularization to maximize the similarity of soft assignments and target distributions. During the past few years, AE is also introduced to multi-view subspace clustering. Sun et al. [45] used self-supervised strategy to improve the unified subspace representation learning. Zhu et al. [69] simultaneously learned a set of view-specific self-expression representations, then which were combined into a common self-expression representation. Wang et al. [50] learned a unified subspace representation from multi-view discriminative feature spaces. Cui et al. [10] proposed the spectral supervisor to guide the learning of consensus subspace representation. The clustering performance of the above deep multi-view subspace clustering approaches are excellent, but their abilities of exploiting the association between multiple subspace representations still need to be improved. For instance, [45, 50]

directly learn the consistent self-expression representation from multi-view latent features refined by AEs, which could not capture the disparate characteristics of varying views, thus failing to utilizing the complementary information. [69] applies a **Hilbert Schmidt Independence Criterion (HSIC)** regularization term to reinforce the diversity of different views, this indistinguishable alienation of different views may render it difficult to obtain the agreement of them. As for [10], a weighted fusion layer is used to integrate all self-expression representations, which does not harness the view correlations in insightful ways. Contrastive learning [18] is an emerging self-supervised strategy that aims to maximize the similarity between positive pairs whereas minimize the similarity between negative pairs. In MVC scenarios, there is a natural contrastive relationship between varying views, thus giving rise to some multi-view contrastive clustering methods [20, 26, 48, 58, 63]. These methods enhance the discrimination of latent representations by maximizing the similarity of positive samples and minimizing the similarity of negative sample pairs from different views, belonging to the feature-level calibration. Nonetheless, an important objective reality in multi-view data are that there may be large modality gap of data under different views, which can drive the distance between instance pairs to be extremely huge, rendering the contrast process difficult. Hence, how to mitigate modal isolation and improve the contrast quality is an vital motivation of this paper. Additionally, most of MVC methods cannot handle the multi-view out-of-samples, which are not involved in the construction of the similarity graph or the training of the clustering network. To cluster the out-of-samples, most clustering algorithms or networks have to be reexecuted or retrained, which is time-consuming and laborious.

We are inspired by the idea of contrastive learning, and propose a **Subspace-Contrastive Multi-View Clustering (SCMC)** method. Specifically, in order to perceive the nonlinear structures in multi-view data, we employ view-specific AEs to encode the initial features into multiple compact spaces, wherein the respective subspace representations are further learned through the self-expression layers, such that the semantic information of data belonging to disparate modalities can be unified into a common semantic space. Thus, we consider the same sample under different views as the positive pairs, and the rest of pairs are considered as negative, Figure 2 illustrates the manner of constructing positive and negative pairs. By pairwisely contrasting multiple subspace representations, we bring the sample affinities of positive pairs closer together and the sample affinities of negative pairs further apart, belonging to the structure-level calibration. This operation enhances the discriminability of each subspace representation and explores the complementary information within them, which is also different from the discrimination-induced regularization achieved by the indistinguishable mutual exclusion between various representations in literatures [50, 69]. To obtain a consistent affinity matrix, we use a weighted fusion scheme to merge multiple subspace representations. Moreover, the graph regularization is applied to encode the local structures inside the learned subspaces, further fine-tuning the suitable affinities between samples. In addition, we further propose an extension mechanism for the multi-view out-of-samples, namely, the **Consistent Sparse Representation (CSR)** learning method over the multi-view in-samples, thus directly achieving the clustering for out-of-samples instead of retraining the clustering network.

Finally, abundant experiments on ten challenging datasets are implemented to verify the effectiveness of the proposed SCMC. The major contributions of this paper are summarized as follows:

- —We nonlinearly map the multi-view data into compact feature subspaces via the AEs, then regard different subspace representations as contrast entities and perform the structure-level contrastive learning, thus exploring the complementarity between heterogeneous views.
- —We obtain an initial unified affinity matrix through the weighted aggregation, to capture the local geometric structures of multiple subspaces, the graph regularization is utilized to further fine-tune the affinities between instances. Thus, the inter-view consistency is well guaranteed.

—We propose an extension mechanism to handle the multi-view out-of-samples, that is, the CSRs of out-of-samples over the in-samples are learned, thus directly obtaining the clustering results for the out-of-samples.
—To demonstrate the validity of the proposed SCMC, we carry out comprehensive experiments on ten multi-view datasets, and the experimental results show that SCMC possesses advanced data clustering capability compared with other MVC methods. Moreover, the proposed CSR is also verified to be effective via the designed experiments.

The rest of this paper are structured as follows. Section 2 briefly reviews the related works. In Section 3, we explicate the proposed SCMC and CSR. In Section 4, we discuss the differences between the existing works and this paper. Experimental details are narrated in Section 5. Finally, the conclusion is summarized in Section 6.

2 Related Works

2.1 Multi-View Subspace Clustering

Multi-view subspace clustering [5] leverages heterogeneous features of data to group samples into a union of diverse subspaces. Self-expression based subspace learning technology has gained widespread attention due to its concise but sound feature characterization capabilities. Some works aimed at exploring a shared subspace representation. For instance, Cai et al. [4] explicitly modeled the consistency and the specificity of multi-view data, and learned a well-structured common affinity matrix. Li et al. [25] proposed a kernel completion schema to learn the compact and low-redundant subspace representations. Wang et al. [53] relied on the information bottleneck theory to discard the superfluous information in raw data, and explored a common subspace representation via the view-common encoder network. Chao et al. [6] leveraged the multiple imputation and ensemble technology to cope with the incomplete multi-view data. Wang et al. [52] adopted the Frobenius norm and $l_{2,1}$-norm to enhance the robustness of consistent graph matrix. For capturing the high-order correlations among views, tensor-oriented methods have been researched. Ji et al. [23] proposed an enhanced tensor nuclear norm to differentiate the contributions of diverse singular values. Qin et al. [42] projected the initial multi-view features into nonlinear subspaces, then captured the high-order correlations via minimizing the rank of representation tensor. Wang et al. [51] considered the existence of data noise in multi-view data and removed it with the help of entropy-regularized tensor learning. To improve the abilities for matching the complex data distributions, some researchers used neural networks to model multi-view data. [50, 69] used deep network frameworks to learn the subspace representations, then they all adopted an exclusive regularization term to boost the complementary information among varying views. Du et al. [11] learned the discriminative multi-view features via adversarial training, based on which the consistent subspace representation was explored. Gao et al. [17] aimed at learning the semantic-invariant representations among multiple heterogeneous features, and sought the optimal cluster divisions by the reinforcement learning.

2.2 Contrastive Learning

Contrastive learning is one of the research hot spots of the self-supervised learning paradigm over recent years, its central concept is to enhance the similarity between positive instance pairs and weaken the similarity between negative instance pairs. In practice, [7, 43, 67] were proposed in computer vision and natural language processing filed, which successively enhanced the discrimination of data representations by means of contrastive learning. Owing to the presented favorable performance, contrastive learning has gained attention and been applied in clustering field. Li et al. [27] simultaneously contrasted the instance-level and cluster-level representations to strength the

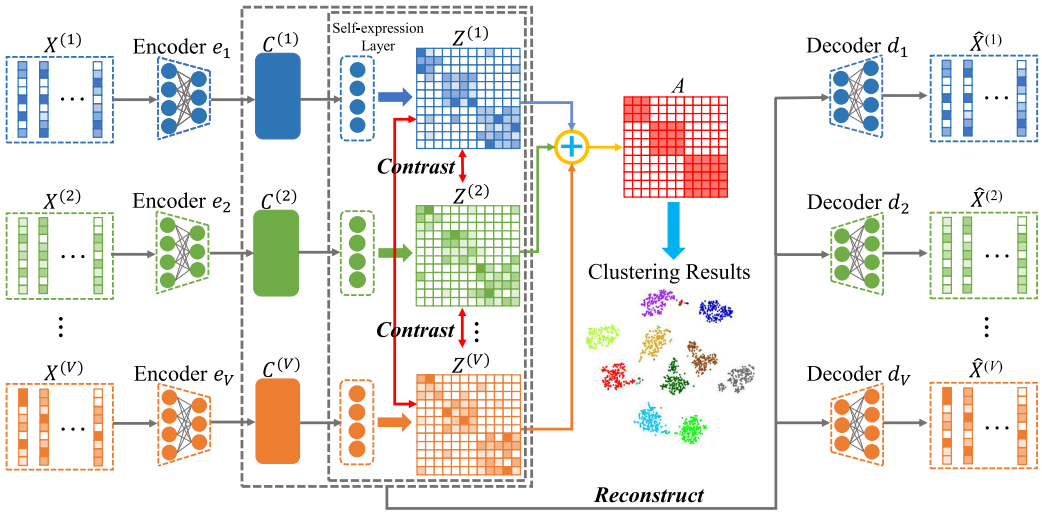


Fig. 1. The framework of the proposed SCMC. To effectively handle high-dimensional and nonlinear structures in multi-view data, we use V view-specific encoders to encode the initial multi-view features $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$ as the compact embedding features $\{\mathbf{C}^{(v)}\}_{v=1}^{V}$. Thus, $\{\mathbf{C}^{(v)^T}\}_{v=1}^{V}$ pass through the multiple self-expression layers to obtain the features $\{\mathbf{C}^{(v)^T}\mathbf{Z}^{(v)}\}_{v=1}^{V}$, which are fed into the V view-specific decoders to reconstruct the recovered data $\{\hat{\mathbf{X}}^{(v)}\}_{v=1}^{V}$. Notably, $\{\mathbf{Z}^{(v)}\}_{v=1}^{V}$ are essentially the coefficient matrices of multiple self-expression layers, also called the subspace representations. We contrast these subspace representations in pairs to exploit the complementary information between them. Further, a weighted fusion of all subspace representations is performed to obtain a unified affinity matrix while the graph regularization is adopted to fine-tune the affinities. Finally, the spectral clustering algorithm is employed to acquire the clustering results.

separability of samples belonging to different clusters. Liu et al. [34] adopted the data preprocessing and multilayer perceptions to achieve the simple contrastive graph clustering, alleviating the burden of high computational complexity. Furthermore, researchers have extended single-view contrastive clustering to multi-view cases, Xu et al. [58] conducted data reconstruction in a low-level space and punished the consistent objectives via a contrastive scheme in a high-level space. Yan et al. [60] proposed a structure-guided contrastive schema to align common and view-specific semantic information. Furthermore, some works also made the progress for combining the contrastive learning and multi-view subspace clustering. Du et al. [12] performed the pairwise contrast between multiple heterogeneous features via binary cross-entropy loss. Cheng et al. [9] contrasted the multi-view features in the latent space and utilized the HSIC regularization to strengthen the diversity of different subspace representations. Zhang et al. [66] proposed an MVC-driven contrast regularizer, which regarded the neighboring nodes of a node as the positive samples as well.

3 The Proposed Method

In this section, we first explain the motivations for proposing the SCMC method. Second, we present the objective functions of the proposed SCMC. Thus, the specific network architectures are summarized, which are also graphically illustrated in Figure 1 for better comprehension. Further, we conduct a series of analyses on the proposed method, including the optimization method, the training details, the extension for the out-of-samples, and the analysis of time complexity.

3.1 Motivation

(1) Existing deep multi-view subspace clustering methods [50, 69] enhance the discrimination of different representations through an exclusive regularization term, but this undifferentiated disparity of all representations may render the models difficult to acquire the agreement across views. Inspired by contrastive learning, we differentially treat samples from various views, construct cross-view positive and negative pairs, and strengthen the discrimination of subspace representations by bringing positive pairs closer and separating negative pairs.
(2) The semantic information gap between different modalities in multi-view data [57] could be very large. For example, the HOG feature of an image describes completely different semantic information from the LBP feature, then the corresponding feature representations are extremely disparate. Most current multi-view contrastive clustering methods [9, 58] are based on feature-level contrast and may suffer from the above drawback. In light of this, we explore the subspace representations of all views to unify the semantic information from heterogeneous views into a joint semantic space, achieving the structure-level contrast.

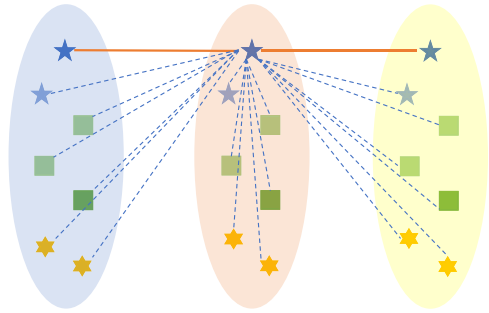
3.2 Objective Function

(1) Reconstruction and Subspace Losses: A multi-view dataset is denoted by $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$. Specifically, $\mathbf{X}^{(v)} \in \mathbb{R}^{N\times d^{(v)}}$ is feature matrix of the vth view, where N and $d^{(v)}$ represent the number of instances and the feature dimension, respectively. For the original feature $\mathbf{X}^{(v)}$, it may be high-dimensional and nonlinear, which poses difficulties for the downstream tasks. Hence, we use multiple view-specific encoders to nonlinearly map $\mathbf{X}^{(v)}$ into a latent low-dimensional space. For the vth view, its encoder is formulated as

$$\mathbf{C}^{(v)} = \mathscr{e}_v(\mathbf{X}^{(v)}|\mathbf{W}_e^{(v)}, \mathbf{b}_e^{(v)}), \tag{1}$$

where $\mathbf{C}^{(v)}$ is the embedding feature after $\mathbf{X}^{(v)}$ passing the encoder $\mathscr{e}_v(\cdot)$, $\mathbf{W}_e^{(v)}$ and $\mathbf{b}_e^{(v)}$ indicate the weight matrix and bias vector in the encoder, respectively. Mathematically, nonlinearity [33, 44] expresses the fact that the dependent variables do not have a linear or direct relationship with the independent variables, and the variation of output is not proportional to the variation of input. In addition, [33, 44] indicate that the output of each layer for a neural network-based model is actually a high-level semantic feature embedding. Under the action of activation functions such as Relu, Sigmoid, and so forth, the data are nonlinearly mapped to a compact representation at each layer. Finally, with the guidance of loss function, the output of last layer is a discriminative representation that is strongly correlated with the downstream tasks.

Current multi-view contrastive clustering tends to first project multi-view data into compact feature spaces and then perform contrast between them. However, the semantic information of data from heterogeneous views can be very disparate. Even if some works such as [58] consider projecting data from different views into a unified feature space through a shared network, the distance between pairs could still be very large. It is difficult to pull together the positive pairs of different modalities. Subspace representation is not only an informative low-dimensional representation form of data, but it also contains an important property, i.e., it portrays the affinity relationship between sample points. For example, given a subspace representation $\mathbf{Z} \in \mathbb{R}^{N\times N}$, $\mathbf{Z}_{ij}$ measures the affinity between the ith instance and the jth instance. Thus, the ith row $\mathbf{Z}_i$ can be considered as a low-dimensional subspace representation of the ith instance, but also as the affinities of the ith instance with other instances. Therefore, there are still differences in subspace representations $\{\mathbf{Z}^{(v)}\}_{v=1}^{V}$ from diverse views, but their semantic information remains consistent, alleviating the dilemma in the contrast



(a) View 1 (b) View 2 (c) View 3

Fig. 2. The diagram of constructing the positive and negative pairs. Let us take three views as an example, the first data point of view 2 and the same data points of the other two views are positive pairs (connected by solid lines). The first data point of view 2 and the remaining data points of the other two views and its own view are negative pairs (connected by dashed lines).

process. We learn the subspace representation $\mathbf{Z}^{(v)}$ of the vth view by

$$\min_{\mathbf{C}^{(v)},\mathbf{Z}^{(v)}} ||\mathbf{C}^{(v)^T} - \mathbf{C}^{(v)^T}\mathbf{Z}^{(v)}||_F^2. \tag{2}$$

In the networks, $\mathbf{Z}^{(v)}$ is coefficient matrix of the learnable self-expression layer, which is achieved via one-layer fully connected layer without the bias part. The original data are projected to the latent space via the encoder, how to ensure that the features in the latent space maintain highly informative is a key issue. In the literature [49], the authors proved that the essence of the reconstruction loss is to the maximize the lower bound of the mutual information between original features and latent features, which guaranteed that the latent features contain as much important information as possible in the original feature space. Inspired by the previous works, we introduce the reconstruction loss to guide the training of encoders, thus obtaining the compact and informative latent features. After the embedding feature $\mathbf{C}^{(v)^T}$ passes through the self-expression layer $\mathbf{Z}^{(v)}$ to obtain $\mathbf{C}^{(v)^T}\mathbf{Z}^{(v)}$, we reconstruct the data via feeding it into the decoder $\mathscr{d}_v(\cdot)$, the decoding process is formulated as

$$\hat{\mathbf{X}}^{(v)} = \mathscr{d}_v(\mathbf{C}^{(v)^T}\mathbf{Z}^{(v)}|\mathbf{W}_d^{(v)}, \mathbf{b}_d^{(v)}), \tag{3}$$

where $\hat{\mathbf{X}}^{(v)}$ denotes the reconstructed data, $\mathbf{W}_d^{(v)}$ and $\mathbf{b}_d^{(v)}$ represent the coefficient matrix and the bias vector of the decoder network, respectively. For V views, the reconstruction loss and subspace representation learning loss are computed by Equations (4) and (5)

$$\mathscr{L}_{Re} = \min_{\hat{\mathbf{X}}^{(v)}} \sum_{v=1}^{V} ||\mathbf{X}^{(v)} - \hat{\mathbf{X}}^{(v)}||_F^2 \tag{4}$$

$$\mathscr{L}_{Sub} = \min_{\mathbf{C}^{(v)},\mathbf{Z}^{(v)}} \sum_{v=1}^{V} ||\mathbf{C}^{(v)^T} - \mathbf{C}^{(v)^T}\mathbf{Z}^{(v)}||_F^2. \tag{5}$$

(2) Contrastive Loss: There is a natural contrast between multiple subspace representations $\{\mathbf{Z}^{(v)}\}_{v=1}^{V}$. To exploit this property, we first construct positive and negative pairs across views. For $\mathbf{Z}_i^{(v)}$, it mutually forms a positive pair with the same instances under different views, while its negative samples contain all the instances except the positive samples. Figure 2 provides a graphical illustration of how positive and negative pairs are constructed. Summarily, there are $V-1$ positive

instances and $V(N-1)$ negative instances for $\mathbf{Z}_i^{(v)}$. Thus, we apply consine distance to measure the similarity between pairs, the mathematical form is expressed as follows

$$\Theta\left(\mathbf{Z}_i^{(v_1)}, \mathbf{Z}_j^{(v_2)}\right) = \frac{\left(\mathbf{Z}_i^{(v_1)}\right)\left(\mathbf{Z}_j^{(v_2)}\right)^\top}{\left\|\mathbf{Z}_i^{(v_1)}\right\|\left\|\mathbf{Z}_j^{(v_2)}\right\|}. \tag{6}$$

Taking one view as an example, to achieve the goal of narrowing the positive pairs and widening the negative pairs, we formulate the problem as

$$\begin{aligned} \ell_v = &- \min_{\mathbf{Z}^{(v)},\mathbf{Z}^{(k)}} \sum_{k=1,k\neq v}^{V} \sum_{i=1}^{N} \\ &\log \frac{\exp\left(\Theta(\mathbf{Z}_i^{(v)}, \mathbf{Z}_i^{(k)})/\tau\right)}{\sum_{j=1}^{N}\left(\exp(\Theta(\mathbf{Z}_i^{(v)}, \mathbf{Z}_j^{(v)})/\tau) + \exp(\Theta(\mathbf{Z}_i^{(v)}, \mathbf{Z}_j^{(k)})/\tau)\right)}, \end{aligned} \tag{7}$$

where τ denotes the temperature parameter. We can observe that the numerator is about the calculation of positive pairs, while the denominator is about the calculation of negative pairs. The contrastable loss of V views is computed by

$$\mathcal{L}_{Con} = \min \frac{1}{NV} \sum_{v=1}^{V} \ell_v. \tag{8}$$

(3) Fusion Loss: For aggregating the complementary information in different views, we integrate multiple subspace representations into a consistent affinity matrix in a weighted fusion manner. Specifically, a set of weight coefficients are assigned to varying views, and they can be adaptively optimized in the back propagation process. The problem is written as

$$\begin{aligned} &\mathbf{A} = \mathcal{f}(\{\mathbf{Z}^{(v)}\}_{v=1}^{V}|\Omega) = \sum_{v=1}^{V} \omega^{(v)}\mathbf{Z}^{(v)} \\ &\text{s.t.} \sum_{v=1}^{V} \omega^{(v)} = 1, \omega^{(v)} \geq 0, \end{aligned} \tag{9}$$

where $\mathcal{f}(\cdot)$ denotes the fusion function, Ω represents the learnable coefficients in the fusion function. In addition, an important assumption of graph embedding theory [39, 59] is that two samples closer to each other in the original space retain this property in the new low-dimensional space. We follow this assumption and consider that two instances similar in subspaces under any view, they should have higher affinity in the unified space. Thus, the following minimum problem can be obtained

$$\begin{aligned} \mathcal{L}_{Fu} &= \min_{\mathbf{Z}^{(v)},\mathbf{A}} \sum_{v} \sum_{i} \sum_{j} ||\mathbf{Z}_i^{(v)} - \mathbf{Z}_j^{(v)}||_2^2 \mathbf{A}_{ij} + \sum_{i} \sum_{j} \mathbf{A}_{ij}^2 \\ &= \sum_{v=1}^{V} Tr(\mathbf{Z}^{(v)}\mathbf{L}_\mathbf{A}\mathbf{Z}^{(v)^T}) + ||\mathbf{A}||_F^2 \\ &\text{s.t. } \mathbf{A}_i\mathbf{1} = 1, \mathbf{A}_{ij} \geq 0, \mathbf{A}_{ii} = 0, \end{aligned} \tag{10}$$

where $\mathbf{A}_{ij}$ is the (i,j)-th element in the uniform affinity matrix $\mathbf{A}$, $\mathbf{A}_i$ denotes the ith row of $\mathbf{A}$, $\mathbf{1}$ is a vector with all elements of 1. The constraints on $\mathbf{A}$ aim to avoid the trivial solutions. One may think that a unified affinity $\mathbf{A}$ can be learned directly through Equation (10), and the weighted fusion mechanism seems to be unnecessary. Our intension is to initially obtain a consistent affinity

matrix $\mathbf{A}$ through Equation (9) to avoid that all elements of $\mathbf{A}$ are zeros, which makes Equation (10) unable to optimize, then we use Equation (10) to further fine-tune the affinities between samples. At present, we give the final objective function of the proposed SCMC, which is written as

$$\begin{aligned}
\mathscr{L} &= \mathscr{L}_{Re} + \gamma_1 \mathscr{L}_{Sub} + \gamma_2 \mathscr{L}_{Con} + \gamma_3 \mathscr{L}_{Fu} \\
&= \sum_{v=1}^{V} \left(||\mathbf{X}^{(v)} - \hat{\mathbf{X}}^{(v)}||_F^2 + \gamma_1 ||\mathbf{C}^{(v)} - \mathbf{C}^{(v)}\mathbf{Z}^{(v)}||_F^2 \right) \\
&- \gamma_2 \sum_{v=1}^{V} \sum_{k=1, k\neq v}^{V} \sum_{i=1}^{N} \\
&log \frac{\exp\left(\Theta(\mathbf{Z}_i^{(v)}, \mathbf{Z}_i^{(k)})/\tau\right)}{\sum_{j=1}^{N} \left(\exp(\Theta(\mathbf{Z}_i^{(v)}, \mathbf{Z}_j^{(v)})/\tau) + \exp(\Theta(\mathbf{Z}_i^{(v)}, \mathbf{Z}_j^{(k)})/\tau)\right)} \\
&+ \gamma_3 \sum_{v=1}^{V} Tr(\mathbf{Z}^{(v)} \mathbf{L}_{\mathbf{A}} \mathbf{Z}^{(v)^T}) + \gamma_3 ||\mathbf{A}||_F^2, \\
&\text{s.t.} \sum_{j=1}^{N} \mathbf{A}_{ij} = 1, \mathbf{A}_{ij} \geq 0, \mathbf{A}_{ii} = 0,
\end{aligned} \tag{11}$$

where γ_1, γ_2, and γ_3 are three nonnegative tradeoff parameters. After the optimization based on back propagation, the consistent affinity $\mathbf{A}$ is obtained, we perform the spectral clustering algorithm on the matrix $(\mathbf{A} + \mathbf{A}^T)/2$ to get the data labels.

We summarize how the complementarity and consistency between views are captured by the proposed SCMC. we first nonlinearly project the original features of each view into the subspaces to unify their semantic meanings. Then, the strategy of contrastive learning is adopted to capture the complementary information between different views. We specify that a sample forms the positive instance pairs with the same samples from other views while forming the negative instance pairs with other samples. Guided by the contrastive learning loss, the positive pairs are approaching while the negative pairs are alienating. Essentially, the view-invariant features are learned via the contrast manner while enhancing the discrimination of the samples in the feature space, thus capturing the complementarity hidden in different views. Further, to explore the consistency across different views, we use the graph regularization to learn the view-shared affinity matrix. Specifically, the graph embedding theory reveals to us that two data points close together in the original space retain the property in the projected subspace. Therefore, we constrain two samples similar in either subspaces should have similarity in the view-shared affinity matrix.

3.3 Network Architecture

Based on the introduction of the objective function above, we sketch the network architecture of the proposed SCMC herein. V three-layer encoders embed multi-view data $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$ into compact features $\{\mathbf{C}^{(v)}\}_{v=1}^{V}$. Next, $\{\mathbf{C}^{(v)^T}\}_{v}^{V}$ pass the V self-expression layers to obtain the matrix $\{\mathbf{C}^{(v)^T}\mathbf{Z}^{(v)}\}_{v=1}^{V}$, respectively. Concretely, the self-expression layer is achieved by a one-layer linear layer discarding the bias part. Then, $\{\mathbf{C}^{(v)^T}\mathbf{Z}^{(v)}\}_{v=1}^{V}$ is fed into V decoders symmetrical to the encoders' structures to decode the reconstructed data $\{\hat{\mathbf{X}}^{(v)}\}_{v=1}^{V}$, respectively. In the encoding and decoding processes, the activation function $Relu(\cdot)$ is adopted. Furthermore, we contrast the learned subspace representations $\{\mathbf{Z}^{(v)}\}_{v=1}^{V}$ with each other, and fuse them into a consistent affinity matrix $\mathbf{A}$ with a group of learnable weights, then the nonnegative $\mathbf{A}$ is obtained via $Relu(\cdot)$ function.

Algorithm 1: SCMC

Input: Multi-view data $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$, parameters γ_1, γ_2, γ_3, and number of clusters c.
Output: Consistent affinity matrix $\mathbf{A}$.

1: Initialize multiple view-specific AEs with the parameters after pre-training, initialize the learning rate to 0.0001, the training epochs to 500.
2: **for** *epoch* = 1 to *training epochs* **do**
3: Compute the reconstruction loss $\mathscr{L}_{Re}$ by Equation (4);
4: Compute the subspace learning loss $\mathscr{L}_{Sub}$ by Equation (5);
5: Compute the contrast loss $\mathscr{L}_{Con}$ by Equation (8);
6: Obtain the initial consistent affinity matrix $\mathbf{A}$ by Equation (9);
7: Compute the local structures loss $\mathscr{L}_{Fu}$ by Equation (10);
8: Compute the overall objective loss $\mathscr{L}$ by Equation (11) and update the network parameters via back propagation;
9: **end for**
10: Performing the spectral clustering algorithm on $\frac{(\mathbf{A}+\mathbf{A}^T)}{2}$ to acquire the data labels.

For fine-tuning the affinities between instances, the graph regularization is leveraged to protect the local structures within multiple subspace representations.

3.4 Optimization of Multiple Variables

Herein, we derive the gradient of each variable in each loss. For simplicity, the bias of linear layer is not considered. Furthermore, the activation function $Relu(\cdot)$ is used, that is, $Relu(x) = x$ if $x > 0$, otherwise $Relu(x) = 0$, so the activation function is not presented in the derivation. We recall the four proposed losses, which are written as

$$\begin{aligned}
\mathscr{L}_{Re} &= ||\mathbf{X}^{(v)} - \hat{\mathbf{X}}^{(v)}||_F^2 \\
\mathscr{L}_{Sub} &= ||\mathbf{C}^{(v)^T} - \mathbf{C}^{(v)^T}\mathbf{Z}^{(v)}||_F^2 \\
\mathscr{L}_{Con} &= -\sum_{k=1,k\neq v}^{V}\sum_{i=1}^{N}\log\frac{\exp\left(\Theta(\mathbf{Z}_i^{(v)},\mathbf{Z}_i^{(k)})/\tau\right)}{\sum_{j=1}^{N}\left(\exp(\Theta(\mathbf{Z}_i^{(v)},\mathbf{Z}_j^{(v)})/\tau)+\exp(\Theta(\mathbf{Z}_i^{(v)},\mathbf{Z}_j^{(k)})/\tau)\right)} \\
\mathscr{L}_{Fu} &= \sum_v\sum_i\sum_j ||\mathbf{Z}_i^{(v)} - \mathbf{Z}_j^{(v)}||_2^2\mathbf{A}_{ij} + ||\mathbf{A}_i||_2^2 + (\mathbf{A}_i\mathbf{1} - 1),
\end{aligned} \tag{12}$$

where $\mathbf{C}^{(v)}$ and $\hat{\mathbf{X}}^{(v)}$ are computed by

$$\begin{aligned}
\mathbf{C}^{(v)} &= \mathbf{X}^{(v)}\mathbf{W}_e^{(v)} \\
\hat{\mathbf{X}}^{(v)} &= \mathbf{Z}^{(v)^T}\mathbf{C}^{(v)}\mathbf{W}_d^{(v)} = \mathbf{Z}^{(v)^T}\mathbf{X}^{(v)}\mathbf{W}_e^{(v)}\mathbf{W}_d^{(v)}.
\end{aligned} \tag{13}$$

For the reconstruction loss $\mathscr{L}_{Re}$, taking the derivative of $\mathbf{C}^{(v)}$, we have

$$\frac{\partial\mathscr{L}_{Re}}{\partial\mathbf{C}_{ij}^{(v)}} = \sum_{k,l}\frac{\partial\mathscr{L}_{Re}}{\partial\hat{\mathbf{X}}_{k,l}^{(v)}}\frac{\partial\hat{\mathbf{X}}_{kl}^{(v)}}{\partial\mathbf{C}_{ij}^{(v)}}. \tag{14}$$

Expanding $\frac{\partial \hat{\mathbf{X}}_{kl}^{(v)}}{\partial \mathbf{C}_{ij}^{(v)}}$ leads to

$$\frac{\partial \hat{\mathbf{X}}_{kl}^{(v)}}{\partial \mathbf{C}_{ij}^{(v)}} = \frac{\partial \sum_{a,b} (\mathbf{Z}_{ka}^{(v)^T} \mathbf{C}_{ab}^{(v)} \mathbf{W}_{d,bl}^{(v)})}{\partial \mathbf{C}_{ij}^{(v)}} = \frac{\partial \mathbf{Z}_{ki}^{(v)^T} \mathbf{C}_{ij}^{(v)} \mathbf{W}_{d,jl}^{(v)}}{\partial \mathbf{C}_{ij}^{(v)}} = \mathbf{Z}_{ki}^{(v)^T} \mathbf{W}_{d,jl}^{(v)} \delta_{lj}, \tag{15}$$

if $l = j$, $\delta_{lj} = 1$, otherwise $\delta_{lj} = 0$. Thus, it has

$$\frac{\partial \mathscr{L}_{Re}}{\partial \mathbf{C}_{ij}^{(v)}} = \sum_{k,l} \frac{\partial \mathscr{L}_{Re}}{\partial \hat{\mathbf{X}}_{kl}^{(v)}} \mathbf{Z}_{ki}^{(v)^T} \mathbf{W}_{d,jl}^{(v)} \delta_{lj} \Rightarrow \frac{\partial \mathscr{L}_{Re}}{\partial \mathbf{C}^{(v)}} = \mathbf{Z}^{(v)} \frac{\partial \mathscr{L}_{Re}}{\partial \hat{\mathbf{X}}^{(v)}} \mathbf{W}_d^{(v)^T}, \tag{16}$$

where $\frac{\partial \mathscr{L}_{Re}}{\partial \hat{\mathbf{X}}^{(v)}} = -2(\mathbf{X}^{(v)} - \hat{\mathbf{X}}^{(v)})$.

Taking the derivative with respect to $\mathbf{Z}^{(v)}$, we have

$$\hat{\mathbf{X}}^{(v)^T} = \mathbf{W}_d^{(v)^T} \mathbf{C}^{(v)^T} \mathbf{Z}^{(v)}. \tag{17}$$

$$\frac{\partial \mathscr{L}_{Re}}{\partial \mathbf{Z}_{ij}^{(v)}} = \sum_{k,l} \frac{\partial \mathscr{L}_{Re}}{\partial \hat{\mathbf{X}}_{kl}^{(v)^T}} \frac{\partial \hat{\mathbf{X}}_{kl}^{(v)^T}}{\partial \mathbf{Z}_{ij}^{(v)}}. \tag{18}$$

Focusing on the $\frac{\partial \hat{\mathbf{X}}_{kl}^{(v)^T}}{\partial \mathbf{Z}_{ij}^{(v)}}$, it can be derived

$$\frac{\partial \hat{\mathbf{X}}_{kl}^{(v)^T}}{\partial \mathbf{Z}_{ij}^{(v)}} = \frac{\partial \sum_a (\mathbf{W}_d^{(v)^T} \mathbf{C}^{(v)^T})_{ka} \mathbf{Z}_{al}^{(v)}}{\partial \mathbf{Z}_{ij}^{(v)}} = \frac{\partial (\mathbf{W}_d^{(v)^T} \mathbf{C}^{(v)^T})_{ki} \mathbf{Z}_{il}^{(v)}}{\partial \mathbf{Z}_{ij}^{(v)}} = \left(\mathbf{W}_d^{(v)^T} \mathbf{C}^{(v)^T}\right)_{ki} \delta_{lj}, \tag{19}$$

if $l = j$, $\delta_{lj} = 1$, otherwise $\delta_{lj} = 0$. Thus, we have

$$\frac{\partial \mathscr{L}_{Re}}{\partial \mathbf{Z}_{ij}^{(v)}} = \sum_{k,l} \frac{\partial \mathscr{L}_{Re}}{\partial \hat{\mathbf{X}}_{kl}^{(v)^T}} (\mathbf{W}_d^{(v)^T} \mathbf{C}^{(v)^T})_{ki} \delta_{lj} \Rightarrow \frac{\partial \mathscr{L}_{Re}}{\partial \mathbf{Z}^{(v)}} = \frac{\partial \mathscr{L}_{Re}}{\partial \hat{\mathbf{X}}^{(v)^T}} \mathbf{C}^{(v)} \mathbf{W}_d^{(v)}. \tag{20}$$

Similar to the process of taking the derivatives for $\mathbf{C}^{(v)}$ and $\mathbf{Z}^{(v)}$, we can obtain

$$\frac{\partial \mathscr{L}_{Re}}{\partial \mathbf{W}_e^{(v)}} = \mathbf{X}^{(v)^T} \mathbf{Z}^{(v)} \frac{\partial \mathscr{L}_{Re}}{\partial \hat{\mathbf{X}}^{(v)}} \mathbf{W}_d^{(v)^T}. \tag{21}$$

$$\frac{\partial \mathscr{L}_{Re}}{\partial \mathbf{W}_d^{(v)}} = \mathbf{W}_e^{(v)^T} \mathbf{X}^{(v)^T} \mathbf{Z}^{(v)} \frac{\partial \mathscr{L}_{Re}}{\partial \hat{\mathbf{X}}^{(v)}}. \tag{22}$$

For the subspace loss $\mathscr{L}_{Sub}$, taking the derivative with respect to $\mathbf{C}^{(v)}$, $\mathbf{Z}^{(v)}$, and $\mathbf{W}_e^{(v)}$, respectively, we have

$$\frac{\partial \mathscr{L}_{Sub}}{\partial \mathbf{C}^{(v)}} = -2\mathbf{Z}^{(v)} (\mathbf{C}^{(v)} - \mathbf{Z}^{(v)^T} \mathbf{C}^{(v)}). \tag{23}$$

$$\frac{\partial \mathscr{L}_{Sub}}{\partial \mathbf{Z}^{(v)}} = -2\mathbf{C}^{(v)} (\mathbf{C}^{(v)^T} - \mathbf{C}^{(v)^T} \mathbf{Z}^{(v)}). \tag{24}$$

$$\frac{\partial \mathscr{L}_{Sub}}{\partial \mathbf{W}_e^{(v)}} = \mathbf{X}^{(v)^T} \frac{\partial \mathscr{L}_{Sub}}{\partial \mathbf{C}^{(v)}}. \tag{25}$$

For the contrastive loss $\mathscr{L}_{Con}$, taking the derivative of $\mathbf{Z}_i^{(v)}$, it can be obtained

$$\frac{\partial \mathscr{L}_{Con}}{\partial \mathbf{Z}_i^{(v)}} = \sum_{k=1,k\neq v}^{V} 1/(t_0 t_1)\mathbf{Z}_i^{(k)^T} - 1/(t_3 t_1)\mathbf{Z}_i^{(k)}\mathbf{Z}_i^{(v)^T}\mathbf{Z}_i^{(v)^T} - \sum_j (A_j \mathbf{Z}_j^{(v)} + B_j \mathbf{Z}_j^{(k)} - C_j \mathbf{Z}_i^{(v)}), \quad (26)$$

where $t_0 = ||\mathbf{Z}_i^{(v)}||_2$, $t_1 = ||\mathbf{Z}_i^{(k)}||_2$, $t_2 = ||\mathbf{Z}_j^{(v)}||_2$, $t_3 = t_0^3$ and A_j, B_j, and C_j are corresponding coefficients.

For the fusion loss $\mathscr{L}_{Fu}$, when taking derivative of $\mathbf{Z}^{(v)}$, we can obtain

$$\frac{\partial \mathscr{L}_{Fu}}{\partial \mathbf{Z}^{(v)}} = 2\sum_{v=1}^{V}(\mathbf{Z}^{(v)}\mathbf{L}_{\mathbf{A}} + \mathbf{Z}^{(v)}\mathbf{L}_{\mathbf{A}}^T). \quad (27)$$

When taking the derivative of $\mathbf{A}_i$, we have

$$\frac{\partial \mathscr{L}_{Fu}}{\partial \mathbf{A}_i} = \sum_v \sum_j ||\mathbf{Z}_i^{(v)} - \mathbf{Z}_j^{(v)}||_2^2 + 2\mathbf{A}_i + \mathbf{1}^T. \quad (28)$$

From the above derivation, we can observe that each variable of each loss has its derivative. Through the iterative optimization, the overall loss becomes decreasing and converges.

3.5 Training Details

The training process of the network is divided in two steps: Pre-training and overall training. First, we pre-train V three-layer AEs to initial their parameters. The purpose of pre-training is to mitigate the difficulties of training the overall network caused by all zeros in the AEs' parameters and the possibility of generating trivial solutions. Second, we train the overall network, the parameters of V AEs, V self-expression layers, a group of learnable weights, and the uniform affinity matrix are iteratively optimized. Adam is adopted as the optimizer, and the learning rate is set to 0.0001. The experiments are run on a server equipped with Intel(R) Core(TM) i9-10980XE, RTX 3090 GPU, and 128G RAM. Algorithm 1 provides a summary of the main steps of the proposed SCMC.

3.6 Extension for the Out-of-Samples

In the practical applications, some samples fail to participate in similarity construction or deep network training. these unseen samples are called out-of-samples [2]. Common clustering methods cannot directly cluster the out-of-samples unless the clustering methods are reexecuted or the network is re-trained, this challenge is equally faced in the MVC. Hence, how to empower the proposed SCMC with the capability of handling the out-of-samples without retraining the network is a matter we consider.

Herein, we propose a CSR construction method over the in-samples for the out-of-samples for bridging the gap. Specifically, the sparsity theory [13, 41] states that any data point can be fitted by a linear combination of a set of basis. Furthermore, assuming that the all samples (including in-samples and out-of-samples) and the in-samples are in independently and identically distributed, thus the subspaces spanned by the in-samples can approximate the subspaces spanned by the all samples. Given the multi-view in-samples $\{\mathbf{X}^{(v)}\}_{v=1}^V$ and the multi-view out-of-samples $\{\hat{\mathbf{X}}^{(v)}\}_{v=1}^V$, where $\mathbf{X}^{(v)} \in \mathbb{R}^{N\times d^{(v)}}$ and $\hat{\mathbf{X}}^{(v)} \in \mathbb{R}^{\hat{N}\times d^{(v)}}$, we first use the proposed SCMC to generate the cluster labels over the multi-view in-samples $\{\mathbf{X}^{(v)}\}_{v=1}^V$. Then, we expect to explore the CSR of multi-view out-of-samples in the subspaces spanned by the multi-view in-samples. For the ith out-of-sample

$\hat{\mathbf{x}}_i^{(v)}$, the objective function of learning the CSR is formulated as

$$\min_{\mathbf{h}_i} \sum_{v=1}^{V} ||\hat{\mathbf{x}}_i^{(v)} - \mathbf{h}_i \mathbf{X}^{(v)}||_2^2 + \lambda ||\mathbf{h}_i||_2^2, \tag{29}$$

where $\mathbf{h}_i$ denotes the CSR for the ith out-of-sample, λ denotes the tradeoff parameter. To obtain the optimal solution of $\mathbf{h}_i$, we take the derivation of $\mathbf{h}_i$ and have

$$\mathbf{h}_i^* = \sum_{v=1}^{V} \hat{\mathbf{x}}_i^{(v)} \mathbf{X}^{(v)^T} \left(\sum_{v=1}^{V} \mathbf{X}^{(v)} \mathbf{X}^{(v)^T} + \lambda \mathbf{I} \right)^{-1}. \tag{30}$$

Then, we compute the distance between the ith out-of-sample and the jth subspace, which is expressed as

$$d_j(\hat{\mathbf{x}}_i) = \sum_{v=1}^{V} ||\hat{\mathbf{x}}_i^{(v)} - \sigma_j(\mathbf{h}_i) \mathbf{X}^{(v)}||_2^2, \tag{31}$$

where the elements of $\sigma_j(\mathbf{h}_i)$ are the entries in $\mathbf{h}_i$ related to the jth subspace. Naturally, the label belonging to the subspace closest to the ith out-of-sample is its label, i.e.,

$$p(\hat{\mathbf{x}}_i) = \arg\min_j (\{d_j(\hat{\mathbf{x}}_i)\}_{j=1}^{c}), \tag{32}$$

where $p(\hat{\mathbf{x}}_i)$ denotes the assigned label for the ith out-of-sample, c is the number of clusters. More importantly, the proposed mechanism of handing the out-of-samples can be applied as a plugin to any of MVC methods.

3.7 Analysis of Time Complexity

In this subsection, we analyze the time complexity of the proposed SCMC. Specifically, the time complexity mainly arises from four aspects, including the encoding-decoding process, the calculation of self-expression layer, the implementation of contrastive loss, and the computation of graph regularization. Suppose the sizes of hidden layers in the vth encoder are $d^{(v)}, h_1, ..., h_l, p$, respectively. Then, the sizes of hidden layers in the vth decoder are in reverse order. First, multi-view features $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$ are passed through the encoders, it takes the $\mathcal{O}((\sum_{v=1}^{V} d^{(v)} h_1 + V(h_1 h_2 + ... + h_l p))N)$ complexity, and so is the decoding. Let $h_{max} = max(\sum_{v=1}^{V} d^{(v)} h_1, V h_1 h_2, ..., V h_l p)$, the encoding complexity is simplified as $\mathcal{O}(h_{max} N)$. Second, the intermediate latent features $\{\mathbf{C}^{(v)}\}_{v=1}^{V}$ are fed into the self-expression layer, which produces the $\mathcal{O}(pN^2)$ complexity. Third, when the contrast of two views is performed, the $\mathcal{O}((N^3 + N(N-1)))$ complexity is consumed. Considering V views, it costs the $\mathcal{O}(V(V-1)(N^3 + N(N-1)))$ complexity. Fourth, computing the graph regularization incurs $\mathcal{O}(VN^3)$ complexity. In summary, the overall time complexity of the proposed SCMC is $\mathcal{O}(h_{max} N + pN^2 + V(V-1)(N^3 + N(N-1)) + VN^3)$.

4 Discussion

At present, there is a lot of works on MVC, some of which are relevant to the proposed SCMC, it is necessary to discuss the differences between them.

—Compared to the traditional shallow multi-view subspace clustering methods [16, 29, 36], we use neural networks to perceive the nonlinear structures in multi-view data, and expect the multi-view data with complex distributional properties to separate well in subspaces. Then, the contrastive learning strategy and graph regularization are leveraged to explore the complementary and consistent information.

Table 1. Statistics of Eight Datasets

Dataset	Views	Samples	Clusters	Features
ALOI	4	1,079	10	64/64/77/13
GRAZ02	6	1,476	4	512/32/256/500/500/680
NUS-WIDE-v1	6	1,600	8	64/144/73/128/225/500
NUS-WIDE-v2	5	2,000	31	65/226/145/74/129
Reuters	5	1,500	6	21,531/24,892/34,251/15,506/11,547
UCI	3	2,000	10	240/76/6
WikipediaArticles	2	693	10	128/10
Youtube	6	2,000	10	2,000/1,024/64/512/64/647
Animals	2	5,000	50	4,096/4,096
Cifar10	2	10,000	10	768/324

—Compared to the deep multi-view subspace clustering methods [10, 45, 50, 69], which either use the feature fusion or exclusion-induced regularizers to mine the complementary information in multi-view data, we adopt the contrastive learning strategy to explore the cross-view complementarity. The main benefit of the introduction of contrastive learning is that it differently treats different sample pairs, bringing the positive sample pairs closer together while pulling the negative sample pairs farther apart, which facilitates both the learning of view-invariant features and the enhancement of the feature discrimination in the latent space.
—Existing multi-view subspace clustering methods [9, 12, 58] with contrastive mechanism perform the contrast in the latent feature space, aiming to align the latent representations of positive sample pairs in different views while alienating the latent representations of negative sample pairs, which belongs to the feature-level calibration. In contrast, we perform the contrast at the self-expression representation level, it is worth noting that the self-expression representation depict the affinity between samples. We bring the sample affinities of positive sample pairs closer under different views while pulling the sample affinities of negative sample pairs away, which belongs to the structure-level calibration.
—The focus of this work [11] is to explore the multi-view and multi-level self-expression representations. To enhance the discrimination of latent features output by the encoder networks, the adversarial training strategy is used to assist the training of encoder networks. Our work aims to bring the same samples closer together and different samples farther apart by contrasting the self-expression representations from different views, thus strengthening the discrimination of self-expression representations.

5 Experiments

5.1 Multi-view Datasets

Table 1 summarizes the statistics of eight test multi-view datasets. Specifically, *ALOI*[1] contains 1,079 images from 10 objects, 4 features are extracted from these images: Haralick texture feature, HSV color histograms, Color similarities, and RGB color histograms. *GRAZ02*[2] is object categorization dataset, which is composed of 1,476 images with 4 kinds of objects. 6 visual features are extracted, including WT, LBP, SIFT feature, pyramid HOG, SURF feature, GIST feature. *NUS-WIDE*[3] is a famous

[1] https://elki-project.github.io/datasets/multi-view
[2] http://www.emt.tugraz.at/pinz/data/GRAZ_02
[3] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide

image database, we select two subsets from it. *NUS-WIDE-v1* contains 1,600 images of 8 categories, 6 views are Color histograms, Edge direction histograms, Block-wise color moments, Bag of words, Color correlograms, WTs, respectively. *NUS-WIDE-v2* consists of 2,000 images from 31 objects, each image is represented from 5 features: Color Histogram, Edge distribution, Color correlation, Color moments, and WT. *Reuters*[4] contains 1,500 documents with 5 languages, each sample is represented as a bag of words that extracted by the TFIDF-based weighting means. *UCI* [1] is comprised of 2,000 handwritten numeric images ranging in [0, 9], three views are FOU feature, PIX feature, and MOR feature, respectively. *WikipediaArticles*[5] is a document dataset organized by editors, which contains 693 short articles with 10 classes and 2 views. *Youtube*[6] is a multi-view video games dataset with 2,000 instances divided into 10 classes. Each entry has 6 features including HOG feature, CH feature, MFCC feature, VS feature, SS feature, and HME feature. *Animals* contains 5,000 samples with 50 classes, two kinds of features are extracted from DECAF and VGG19 networks, respectively. *Cifar10*[7] is famous image dataset consisted of 10 categories of objects, 10,000 samples is randomly selected, and each item has two features: CH feature and HOG feature. Following the most of deep clustering works [9, 11, 50, 53, 69], we use the all samples in a dataset as the training set and also use all samples as the test set, i.e., the training set is the test set. When the training process using the training samples is completed, we adopt the spectral clustering algorithm over the learned consistent affinity matrix to generate the predicted labels of training samples. Finally, the values of evaluation metrics are computed through comparing the ground truth and predicted labels.

In addition, we explain what complementarity and consistency are in the test datasets. In [57, 68], the authors clarified the meaning of complementarity and consistency in multi-view data. The feature representations from different views characterize different aspects of a same object, one view possesses some unique characteristics that other views do not possess. *The data information between multiple views is complementary with respect to the complete data information.* Although the characteristics of different views are diverse, *they all agree on a consistent latent feature space.* In other words, the feature representations of different views can be regarded as the specific projections from the consistent latent representations. Figure 3 illustrates the ideal stated above. For each multi-view dataset, its complementarity is reflected in the feature representations from multiple domains, taking the ALOI dataset as an example, four kinds of feature representations from four domains consist of the dataset, including Haralick texture feature, HSV color histograms, Color similarities, and RGB color histograms. Its consistency is reflected by the consistent affinity matrix learned via the proposed SCMC.

5.2 Baseline and Compared Multi-View Methods

To illustrate the validity of the proposed SCMC, we select the k-means algorithm as the baseline method. In practice, multiple features are concatenated together to form a unified feature representation, then it is fed into k-means to obtain the clustering labels. In addition, we collect thirteen state-of-the-art MVC approaches as the compared methods, which are briefly introduced as follows.

— *AMGL* [37] proposed a auto-weighted multi-view fusion mechanism, then solved the consensus spectral embedding.
— *SwMC* [38] proposed a self-weighted fusion method of multiple graphs, obtaining the consistent graph with exact connection components.

[4] https://archive.ics.uci.edu/ml/datasets.html
[5] http://lig-membres.imag.fr/grimal/data.html
[6] http://archive.ics.uci.edu/ml/datasets
[7] https://www.cs.toronto.edu/kriz/cifar.html

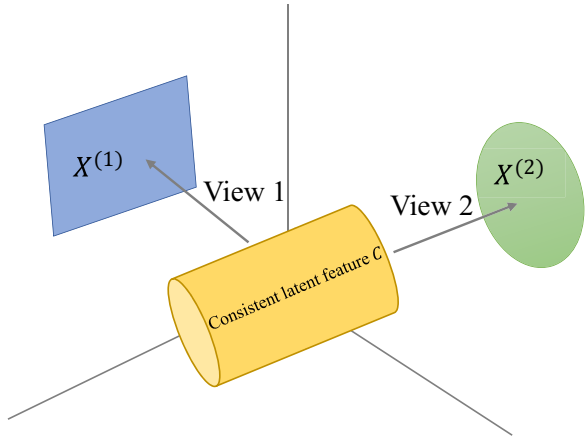
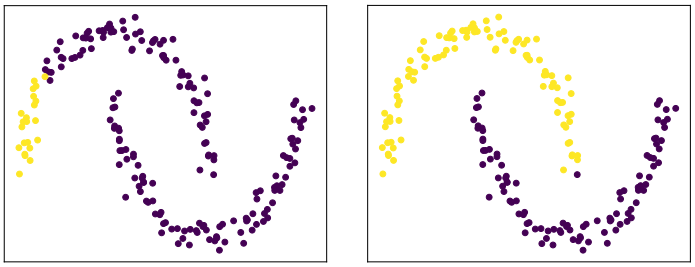


Fig. 3. Illustration of multi-view data. The feature representations $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ of View 1 and 2 depict different aspects for the same object, and they can be regarded as the feature representations obtained from a consistent latent feature through different projections.

- —*TBGL* [55] learned a low-rank bipartite graph tensor representation, then acquired the unified graph via auto-weighted integration.
- —*CSMSC* [36] learned the consensus subspace representation while the specific representations of different views were also explored.
- —*MCGC* [65] designed a disagreement cost function to reinforce the consensus among different graphs.
- —*SM^{2}SC* [64] pursued the view consistency via a variable splitting and a multiplicative decomposition module.
- —*MvDSCN* [69] proposed the diversity and uniformity networks to capture the view-specific and consistent information, respectively.
- —*LMVSC* [24] used anchor graph embedding to fit the global affinity matrix, thus resulting in linear computational complexity.
- —*CGL* [28] optimized the spectral embedding matrices in a low-rank tensor space.
- —*DMSC-UDL* [50] learned a unified subspace representation while enhanced the discrimination between diverse views.
- —*EOMSC-CA* [32] fused the anchor graph scheme and graph construction into a uniform model.
- —*CoMSC* [31] employed the eigendecomposition to obtain the robust representations of multi-view data, from which the consistent self-expressive representation was explored.
- —*MFLVC* [58] proposed a novel contrastive MVC framework that simultaneously learned low-level and high-level features from multi-view data.

CSMSC, *SM^{2}SC*, *LMVSC*, *EOMSC-CA*, and *CoMSC* are shallow models based on subspace learning. *MvDSCN* and *DMSC-UDL* are deep subspace models. *MCGC* is a graph based model. *CGL* is a low-rank tensor learning based model. *MFLVC* is a deep model using contrast strategy. We run the codes published by the authors and set the parameters according to the intervals suggested in these papers.

In the proposed SCMC, we use two kinds of three-layer AEs to encode the multi-view data. The dimensions of each layer of one AE are $[d^{(v)}, 500]$, $[500, 200]$, and $[200, c]$, respectively. The dimensions of each layer of another AE are $[d^{(v)}, 200]$, $[200, 100]$, and $[100, c]$, respectively. $d^{(v)}$ and c denote the feature dimension of the vth view's data matrix and the number of clusters. The



(a) AMGL (b) SCMC

Fig. 4. The visualizations of AMGL and SCMC on Moon dataset.

temperature parameter τ is fixed to 0.1. γ_1, γ_2 are tuned in $\{500, 1,000\}$, $\{0.01, 0.025, 0.03, 0.06, 0.3, 0.4\}$, respectively. γ_3 is set to 0.01. Herein, we illustrate how the selected values of γ_1, γ_2, and γ_3 are obtained. To seek the best parameters, the two-stage grid search strategy is adopted. Specifically, in the first stage, namely, the coarse-grained stage, we empirically set the values of γ_1, γ_2, and γ_3 in $\{0.001, 0.01, 0.1, 1, 10, 100, 1,000\}$, and perform the proposed SCMC on the small-size dataset WikipediaArticles to find the general parameter intervals that can yield the acceptable clustering results. After the coarse-grained stage, we locate the appropriate values of γ_1, γ_2, and γ_3 in $\{100, 1,000\}$, $\{0.01, 1\}$, $\{0.01\}$, respectively. To cope with different datasets having different characteristics, we carry out the fine-grained grid search in the second stage on each dataset. For instance, we tune the value of γ_1 in [100, 1, 000] with a step of 100, tune the value of γ_2 in [0.01, 1] with a step of 0.01.

5.3 Evaluation Metrics

Seven dominant clustering evaluation metrics are adopted to quantify the clustering performance, which are **Accuracy (ACC)**, **Normalized Mutual Information (NMI)**, **Purity, Adjusted Rand Index (ARI)**, F-score, Precision, and Recall respectively. In view of the algorithm stability, each experiment is run ten times, then the mean values are reported. The ranges of ACC, NMI, Purity, F-score, Precision, and Recall are all [0, 1], while the range of ARI is $[-1, 1]$. For all metrics, higher values correspond to better clustering effects. The details of the evaluation metrics are introduced in Appendices A.

5.4 Experimental Results

To verifying the ability of capturing the nonlinear structures of SCMC, we first perform the traditional method AMGL and the proposed SCMC on the Moon dataset, which is a typical nonlinear multi-view dataset with 200 samples and 2 views, and visualize the clustering results in Figure 4. It can be seen that AMGL achieves the inferior performance, misclassifying almost of upper part of the dataset. While SCMC obtains the nearly perfect results, demonstrating the its superior ability of capturing the nonlinear structures. The clustering performance of each view is reported in Tables 20–22 of Appendices B. Tables 2–11 provide the numerical results of all experiments, the best results are bolded and the second best results are underlined. These results reveal several interesting phenomenons and they are explained below. From a holistic perspective, the proposed SCMC outperforms other single-view and MVC methods, which shows that the strategies adopted by the model to improve the representation learning ability is efficient. Especially, on the UCI dataset, SCMC raises the clustering performance by 9%, 2.28%, 8.45%, 9.7%, 8.68%, 14.85%, and 1.47% compared to the suboptimal results. MFLVC also leverages the contrast mechanism to improve the information capability of different representations, whose clustering performance achieves favorable status on several datasets such as GRAZ02, Youtube. While SCMC still achieves better

Table 2. Comparison of Clustering Results (%) on ALOI Dataset

Datasets	Methods	ACC	NMI	Purity	ARI	F-score	Precision
ALOI	K-means	47.49	47.34	48.58	32.98	41.04	33.97
	AMGL	52.28	54.85	55.70	33.18	41.03	34.71
	SwMC	67.10	64.99	69.42	37.29	45.71	33.77
	TBGL	58.29	56.15	61.35	26.24	36.74	25.59
	CSMSC	75.66	73.32	76.68	63.61	67.42	63.80
	MCGC	83.32	74.77	83.32	65.51	69.06	66.61
	SM^{2}SC	23.73	26.34	28.64	13.86	25.29	18.90
	MvDSCN	82.39	82.08	74.88	74.56	79.56	77.87
	LMVSC	73.31	68.62	73.59	58.50	62.89	58.77
	CGL	<u>94.89</u>	<u>91.54</u>	<u>94.89</u>	<u>89.18</u>	<u>90.25</u>	<u>90.03</u>
	DMSC-UDL	79.61	80.16	72.93	67.33	77.71	76.22
	EOMSC-CA	65.80	76.36	66.27	47.24	54.49	39.33
	CoMSC	80.63	81.31	84.62	71.96	74.95	69.63
	MFLVC	82.63	78.57	72.85	69.59	73.89	73.84
	SCMC	**95.74**	**92.72**	**95.74**	**90.91**	**92.16**	**92.18**

The best results are bolded, and the second-best results are underlined.

Table 3. Comparison of Clustering Results (%) on GRAZ02 Dataset

Dataset	Methods	ACC	NMI	Purity	ARI	F-score	Precision
GRAZ02	K-means	35.91	3.20	35.91	3.56	33.83	27.19
	AMGL	46.61	12.70	46.61	10.61	33.94	32.81
	SwMC	38.89	5.64	40.31	6.82	35.10	29.18
	TBGL	46.54	12.56	46.68	11.16	40.23	29.18
	CSMSC	-	-	-	-	-	-
	MCGC	43.02	6.92	43.02	6.44	30.10	30.03
	SM^{2}SC	47.15	12.49	47.76	11.91	34.22	34.09
	MvDSCN	40.44	6.81	50.54	5.93	31.24	29.97
	LMVSC	44.24	8.21	44.24	8.09	31.40	31.23
	CGL	46.46	12.54	46.46	11.43	33.87	33.72
	DMSC-UDL	41.32	8.33	52.24	8.28	32.65	31.05
	EOMSC-CA	42.48	12.19	46.95	10.66	33.33	33.14
	CoMSC	40.79	8.29	43.50	7.82	31.17	31.04
	MFLVC	<u>47.97</u>	<u>13.76</u>	<u>57.18</u>	<u>13.79</u>	<u>35.64</u>	<u>35.35</u>
	SCMC	**51.90**	**16.16**	**59.55**	**14.11**	**37.34**	**37.72**

The best results are bolded, and the second-best results are underlined.

performance, this situation may be attributed to the fact that SCMC applies the subspace learning technology to standardizes data from various modalities into a common semantic space, thus alleviating the difficulties in the contrast process. MvDSCN and DMSC-UDL are two deep multi-view subspace clustering models, from all experimental statistics, their clustering outcomes are not stable. For example, their clustering effects on UCI dataset are relatively superior, and both exceed the baseline method k-means. Nonetheless, on WikipediaArticles dataset, they both produce inferior clustering results than k-means. An interesting observation is that although CGL is a conventional shallow model, it yields promising clustering results, probably because it captures the high-order correlations between views by virtue of the low-rank tensor learning. Furthermore, we present the running time of all compared methods on the test time in the Figure 7. It can be seen that due to the network training consuming time, the proposed SCMC needs relatively longer time

Table 4. Comparison of Clustering Results (%) on NUS-WIDE-v1 Datasets

Dataset	Methods	ACC	NMI	Purity	ARI	F-score	Precision
NUS-WIDE-v1	*K*-means	25.44	14.88	28.69	6.79	23.71	16.26
	AMGL	26.13	16.83	29.63	9.09	26.07	17.37
	SwMC	33.88	17.28	33.88	11.23	25.63	19.67
	TBGL	27.19	14.43	28.63	6.40	24.33	15.81
	CSMSC	34.31	19.58	*38.38*	*13.29*	24.46	23.63
	MCGC	31.75	18.87	35.63	11.35	23.40	21.42
	SM^{2}SC	32.31	19.87	36.81	12.66	23.83	23.18
	MvDSCN	31.13	16.76	35.19	10.01	23.66	23.72
	LMVSC	31.88	*20.63*	36.19	12.38	24.34	22.19
	CGL	31.35	20.36	36.91	12.57	24.52	22.31
	DMSC-UDL	27.19	12.74	32.88	6.91	23.88	19.40
	EOMSC-CA	32.94	20.21	34.00	11.84	*27.11*	19.48
	CoMSC	28.63	11.82	29.88	9.27	20.77	20.40
	MFLVC	<u>34.81</u>	18.68	36.13	**14.33**	26.55	<u>24.62</u>
	SCMC	**36.56**	**21.83**	**40.06**	**14.33**	**27.92**	**28.15**

The best results are bolded, and the second-best results are underlined.

Table 5. Comparison of Clustering Results (%) on four NUS-WIDE-v2 Dataset

Dataset	Methods	ACC	NMI	Purity	ARI	F-score	Precision
NUS-WIDE-v2	*K*-means	15.15	*19.62*	25.55	4.91	9.88	11.16
	AMGL	15.90	17.63	23.50	3.34	8.83	9.18
	SwMC	15.85	9.18	18.35	0.91	11.56	6.36
	TBGL	14.65	4.13	15.40	0.01	11.07	5.92
	CSMSC	13.60	15.89	21.90	3.08	7.55	9.68
	MCGC	15.05	15.92	21.60	3.27	9.20	8.87
	SM^{2}SC	15.15	17.16	24.75	4.46	8.49	11.89
	MvDSCN	14.55	17.90	*25.80*	3.71	10.05	*12.87*
	LMVSC	15.40	18.13	24.35	4.66	8.76	12.04
	CGL	14.37	17.53	23.85	3.79	7.96	10.86
	DMSC-UDL	15.85	14.55	20.35	1.69	**14.26**	10.84
	EOMSC-CA	15.05	13.33	18.90	3.92	11.17	6.10
	CoMSC	12.85	14.39	20.60	2.39	6.69	8.98
	MFLVC	<u>16.35</u>	14.16	23.15	<u>5.35</u>	<u>13.31</u>	10.45
	SCMC	**17.85**	**21.23**	**30.30**	**5.92**	11.97	**15.10**

The best results are bolded, and the second-best results are underlined.

on the small datasets, while the running time is more satisfactory on the large datasets than that of some compared methods such as CSMSC and TBGL.

5.5 Experimental Visualization

We undertake some visualization experiments to directly compare the divergence in representation learning abilities of varying approaches. The t-SNE technology is used to downscale the concatenated UCI dataset to a two-dimensional planes. Figure 5 visualizes the clustering results of ten MVC methods on the UCI dataset, and data points belonging to different clusters are painted in different colors. It can be seen that the ten clusters divided by SCMC rarely involve the samples of other clusters, the segmentation is nearly perfect. On the contrary, the clusters produced by EOMSC-CA are heavily mixed with samples of the other clusters, and even cannot divide enough ten clusters.

Figure 6 presents the visualizations of consistent affinity representations learned by seven compared MVC methods and the proposed SCMC. We can observe that SM^{2}SC and DMSC-UDL

Table 6. Comparison of Clustering Results (%) on Reuters Dataset

Datasets	Methods	ACC	NMI	Purity	ARI	F-score	Precision
Reuters	K-means	42.47	22.09	50.67	16.19	36.63	32.40
	AMGL	42.20	13.19	42.87	5.94	36.72	24.21
	SwMC	44.93	28.49	52.53	10.15	36.87	26.74
	TBGL	32.07	6.27	34.27	0.84	35.02	21.92
	CSMSC	<u>50.27</u>	<u>30.12</u>	<u>58.67</u>	**24.83**	42.06	39.58
	MCGC	45.40	19.06	51.27	17.66	39.37	32.21
	SM^{2}SC	47.93	26.06	52.40	<u>24.76</u>	41.25	40.59
	MvDSCN	49.20	28.59	50.67	19.71	42.46	41.46
	LMVSC	47.40	26.80	51.73	19.64	41.27	33.07
	CGL	44.59	21.21	48.97	20.83	37.35	38.61
	DMSC-UDL	34.93	16.50	43.00	1.85	<u>44.42</u>	30.66
	EOMSC-CA	37.60	12.53	46.40	12.32	30.20	31.95
	CoMSC	-	-	-	-	-	-
	MFLVC	43.40	29.76	**60.53**	24.70	38.13	<u>44.23</u>
	SCMC	**51.80**	**34.47**	58.13	21.83	**50.97**	**44.93**

The best results are bolded, and the second-best results are underlined.

Table 7. Comparison of Clustering Results (%) on UCI Dataset

Datasets	Methods	ACC	NMI	Purity	ARI	F-score	Precision
UCI	K-means	38.76	46.64	44.23	31.35	38.86	35.39
	AMGL	76.28	78.30	78.02	68.69	72.04	67.55
	SwMC	72.95	79.09	75.55	67.17	70.80	63.85
	TBGL	81.40	84.06	83.40	76.13	78.71	72.50
	CSMSC	87.20	77.35	87.20	74.32	76.90	76.24
	MCGC	80.20	79.74	83.55	73.90	76.63	72.87
	SM^{2}SC	84.20	79.94	84.20	74.86	77.37	77.18
	MvDSCN	81.85	71.72	73.05	65.82	69.47	69.31
	LMVSC	<u>87.75</u>	80.55	87.75	76.58	78.93	78.27
	CGL	84.25	<u>90.52</u>	<u>88.55</u>	<u>83.22</u>	<u>85.02</u>	<u>78.86</u>
	DMSC-UDL	78.25	76.66	81.80	66.54	74.75	72.62
	EOMSC-CA	54.80	67.09	55.10	46.28	53.57	39.29
	CoMSC	77.80	78.41	81.55	69.28	72.60	66.85
	MFLVC	79.95	78.36	79.95	69.73	73.31	72.51
	SCMC	**96.75**	**92.84**	**97.00**	**92.92**	**93.70**	**93.71**

The best results are bolded, and the second-best results are underlined.

almost fail to engrave the diagonal-block structures, CGL can clearly highlight important structures on the diagonal, but the boundaries between diagonal-blocks are not apparent. Fortunately, SCMC achieves the clear depiction of diagonal-blocks. In addition, the view-specific and consistent affinity representations produced by SCMC are also visualized in Figure 8. As Figure 8 shows, the abilities to characterize the similarities between instances of the affinity matrices of four single views are not good, Figure 8(e) is generated via averaging all subspace representations, namely $\mathbf{Z}$-sum $= (\mathbf{Z}^{(1)}+\mathbf{Z}^{(2)}+\mathbf{Z}^{(3)}+\mathbf{Z}^{(4)})/4$. Though $\mathbf{Z}$-sum can also clearly portray the diagonal-blocks structures, some non-diagonal elements are also easily observed. Its suppression of non-diagonal affinities is insufficient, which means that it does not protect the local geometric structures of data well.

To evaluate the convergence of the proposed SCMC, we plot the curves of the objective function values and metric values as the number of training epochs increases on eight datasets in Figure 9. It

Table 8. Comparison of Clustering Results (%) on WikipediaArticles Dataset

Datasets	Methods	ACC	NMI	Purity	ARI	F-score	Precision
WikipediaArticles	K-means	54.69	51.48	58.59	39.02	45.80	44.97
	AMGL	55.99	52.69	60.89	39.31	46.10	45.03
	SwMC	51.08	44.81	53.68	19.94	31.49	23.84
	TBGL	29.44	21.40	32.61	6.04	20.80	14.46
	CSMSC	52.03	46.47	55.70	38.36	44.91	46.16
	MCGC	54.40	40.79	56.85	31.68	38.91	40.13
	SM^2SC	55.12	50.83	59.31	40.67	46.95	48.46
	MvDSCN	39.97	32.09	47.04	21.41	31.87	32.62
	LMVSC	55.56	47.46	57.00	33.12	41.00	37.99
	CGL	54.16	49.83	59.42	37.09	44.11	43.21
	DMSC-UDL	44.30	35.01	47.33	22.30	35.53	36.15
	EOMSC-CA	<u>56.13</u>	<u>52.91</u>	<u>61.04</u>	<u>42.30</u>	<u>48.47</u>	<u>49.56</u>
	CoMSC	21.07	7.18	23.23	2.90	13.13	13.61
	MFLVC	43.15	31.54	47.61	24.97	34.09	32.61
	SCMC	**57.86**	**53.74**	**63.20**	**42.54**	**50.61**	**51.01**

The best results are bolded, and the second-best results are underlined.

Table 9. Comparison of Clustering Results (%) on Youtube Dataset

Datasets	Methods	ACC	NMI	Purity	ARI	F-score	Precision
Youtube	K-means	24.56	15.16	27.85	8.19	19.48	15.78
	AMGL	27.65	16.68	27.80	11.47	24.23	16.74
	SwMC	23.00	11.18	24.25	6.27	17.54	14.52
	TBGL	20.80	10.55	21.10	3.72	20.49	11.71
	CSMSC	-	-	-	-	-	-
	MCGC	28.35	13.66	29.70	8.34	17.62	17.33
	SM^2SC	30.58	18.12	34.08	11.14	20.11	19.83
	MvDSCN	30.55	17.67	34.90	10.76	22.77	20.41
	LMVSC	29.15	18.01	33.30	10.06	19.50	18.52
	CGL	<u>32.95</u>	19.65	35.39	12.52	21.48	20.88
	DMSC-UDL	29.95	17.13	35.30	9.39	21.54	19.35
	EOMSC-CA	32.20	18.17	32.60	12.82	23.49	19.11
	CoMSC	24.60	10.41	26.00	5.87	15.42	15.13
	MFLVC	31.80	<u>24.37</u>	<u>38.00</u>	<u>15.98</u>	<u>26.01</u>	<u>22.22</u>
	SCMC	**37.90**	**26.12**	**41.15**	**18.53**	**28.50**	**25.61**

The best results are bolded, and the second-best results are underlined.

can be observed that the objective function values decrease quickly and the metric values increase quickly, though there are fluctuations in the middle of training process, they eventually tend to stabilize.

5.6 Ablation Study

The proposed SCMC consists of four loss components, i.e., $\mathscr{L}_{Re}$, $\mathscr{L}_{Sub}$, $\mathscr{L}_{Con}$, and $\mathscr{L}_{Fu}$. We implement experiments to verify the role played by each loss component. Tables 12, 13, and 14 report the ablation results. When SCMC has only reconstruction loss $\mathscr{L}_{Re}$, we obtain the consensus features by averaging multiple embedding features, then they are fed into the k-means algorithm to yield the clustering results. However, the results are inferior, which means that the latent representations embedded only through AEs are not yet well discriminated. When the subspace loss is introduced, i.e., $\mathscr{L}_{Re} + \mathscr{L}_{Sub}$, the performance is somewhat improved. It is worth noting that once the contrast

Table 10. Comparison of Clustering Results (%) on Animals Dataset

Datasets	Methods	ACC	NMI	Purity	ARI	F-score	Precision
Animals	K-means	30.50	39.78	34.52	17.52	19.33	20.64
	AMGL	43.64	54.43	48.74	18.11	21.02	13.88
	SwMC	54.04	62.47	60.80	14.24	17.63	10.45
	TBGL	34.86	37.73	36.64	4.62	8.85	4.69
	CSMSC	51.42	61.74	56.02	41.92	43.23	45.06
	MCGC	53.44	62.39	58.96	19.59	22.67	13.90
	SM^{2}SC	55.84	66.38	63.78	44.56	45.79	48.63
	MvDSCN	55.18	65.99	60.62	44.38	45.23	48.36
	LMVSC	37.90	57.15	47.54	28.59	30.22	31.18
	CGL	42.57	53.82	47.75	28.96	30.73	29.01
	DMSC-UDL	49.16	62.45	55.06	22.41	47.20	43.35
	EOMSC-CA	54.88	67.21	59.72	**47.54**	42.94	**53.24**
	CoMSC	32.50	40.80	37.96	17.30	19.16	20.02
	MFLVC	28.10	41.45	29.56	18.99	21.42	18.02
	SCMC	**57.06**	**68.82**	**64.16**	45.49	**48.30**	51.57

The best results are bolded, and the second-best results are underlined.

Table 11. Comparison of Clustering Results (%) on Cifar10 Dataset

Datasets	Methods	ACC	NMI	Purity	ARI	F-score	Precision
Cifar10	K-means	15.84	3.81	16.89	1.71	13.80	11.22
	AMGL	10.74	0.37	10.82	0.00	18.17	10.00
	SwMC	17.04	6.05	17.34	1.97	**18.97**	10.94
	TBGL	10.59	0.15	10.60	0.00	18.18	10.00
	CSMSC	17.30	4.26	17.79	2.16	12.23	11.88
	MCGC	18.12	6.17	18.57	3.01	18.96	11.53
	SM^{2}SC	16.90	4.10	17.50	2.05	12.26	11.77
	MvDSCN	17.68	6.65	25.40	2.62	16.07	13.04
	LMVSC	18.55	8.21	19.51	4.05	13.91	13.54
	CGL	19.03	6.74	19.84	3.67	14.46	12.91
	DMSC-UDL	18.76	7.46	23.05	3.39	15.96	13.39
	EOMSC-CA	18.36	7.90	18.43	3.38	10.95	13.18
	CoMSC	17.15	4.09	18.29	2.11	12.60	11.76
	MFLVC	19.62	6.68	24.53	3.26	13.15	13.19
	SCMC	**21.22**	**10.26**	**25.56**	**5.19**	16.75	**14.45**

The best results are bolded, and the second-best results are underlined.

loss is included, the improvements of clustering effects are significant, illustrating that the contrast strategy dose help to augment the discrimination of subspace representations. In general, on the basis of $\mathscr{L}_{Re} + \mathscr{L}_{Sub}$, SCMC performs better with the fusion loss $\mathscr{L}_{Fu}$ than with contrast loss $\mathscr{L}_{Con}$, which indicates that the contribution of $\mathscr{L}_{Fu}$ is greater. Nevertheless, $\mathscr{L}_{Con}$ is essential if the optimal performance is to be achieved. To verify the necessity of weighted fusion manner, we use the method of averaging multiple subspace representations for the ablation experiments, that is, the initial consistent affinity matrix is obtained by $\mathbf{A} = \sum_{v=1}^{V} \mathbf{Z}^{(v)}/V$. The experimental results are showed in the Figure 10, we can see that the clustering results with weight fusion manner outperforms that of average fusion approach, showing that integrating multiple subspace representations with weight assignment is more beneficial for utilizing the complementary information.

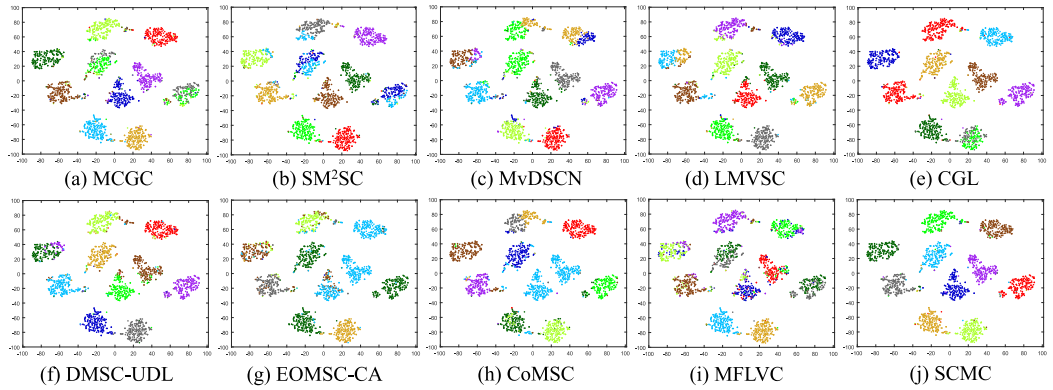


Fig. 5. Visualization of clustering results of ten MVC methods on UCI dataset via t-SNE technology.

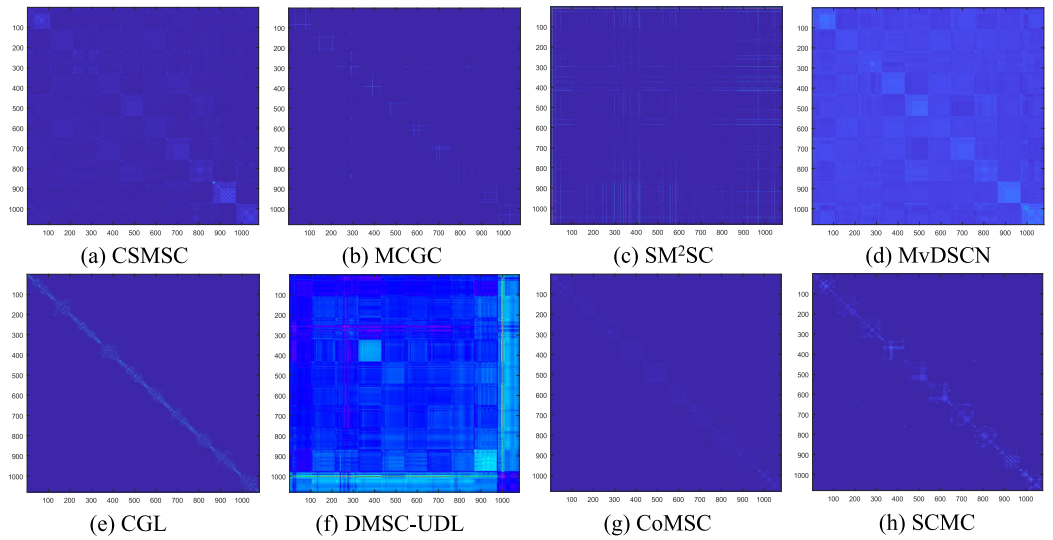


Fig. 6. Visualizations of consistent affinity representations learned via eight MVC approaches on ALOI dataset.

5.7 Investigation of Subspace-Contrastive Effects

We argue that learning subspace representations of multi-view data unifies the semantic information between different modalities, thus facilitating the contrastive training of positive and negative instance pairs. Some comparative experiments are designed to validate the subspace-contrastive effects.

Firstly, we directly perform the contrast between latent representations learned by multiple view-specific AEs, namely removing the subspace loss $\mathscr{L}_{Sub}$. The reconstruction loss $\mathscr{L}_{Re}$ is transformed as

$$\mathscr{L}_{Re} = \sum_{v=1}^{V} ||\mathbf{X}^{(v)} - d_v(e_v(\mathbf{X}^{(v)}|\mathbf{W}_e^{(v)}, \mathbf{b}_e^{(v)})|\mathbf{W}_d^{(v)}, \mathbf{b}_d^{(v)})||_F^2. \tag{33}$$

Meanwhile, the fusion loss $\mathscr{L}_{Fu}$, i.e., Equation (10) is modified as the following form:

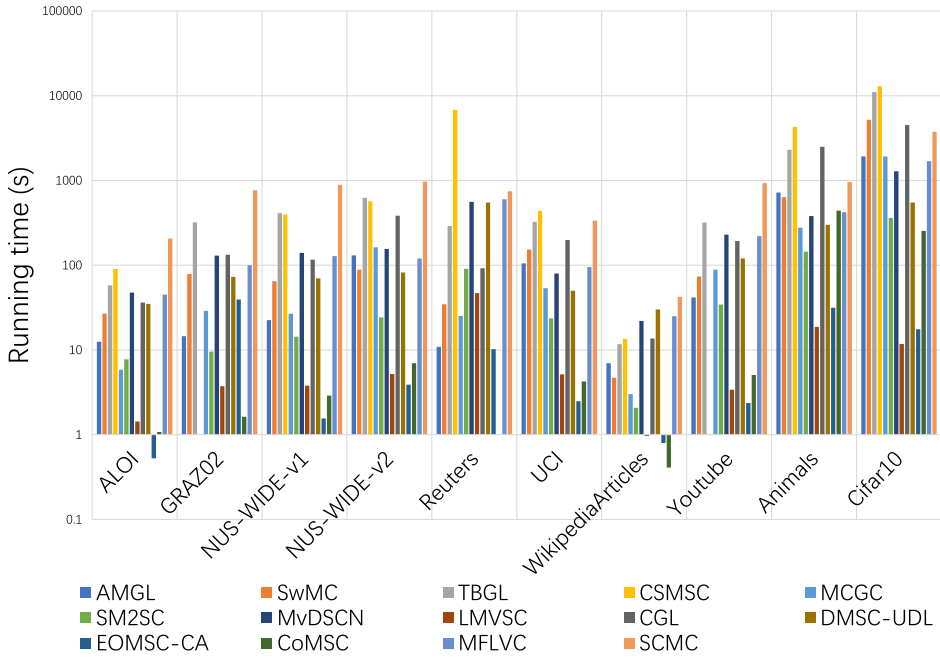


Fig. 7. Comparison of execution time of all compared MVC methods on the test datasets. Considering the large gaps between some values of running time, we perform the $log(\cdot)$ with the base of 10.

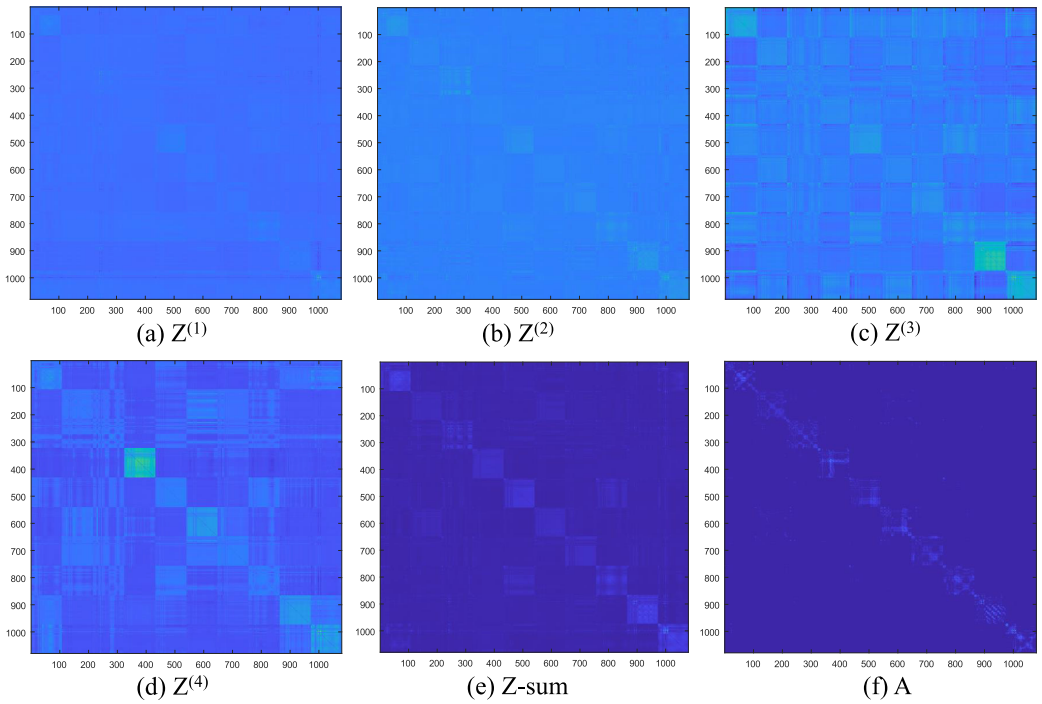
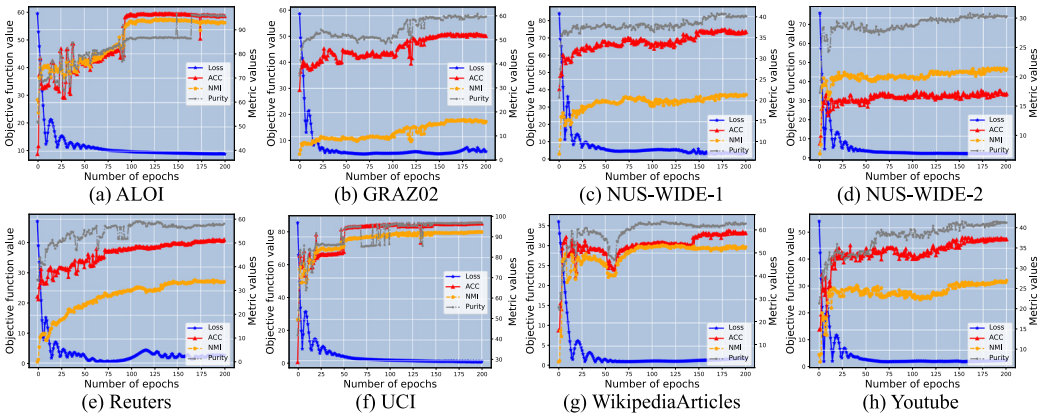


Fig. 8. Comparison of view-specific and consistent affinity matrices generated by the proposed SCMC on ALOI dataset.



(a) ALOI (b) GRAZ02 (c) NUS-WIDE-1 (d) NUS-WIDE-2

(e) Reuters (f) UCI (g) WikipediaArticles (h) Youtube

Fig. 9. The convergence curves of the proposed SCMC on eight datasets.

Table 12. Ablation Results of the Proposed SCMC on Four Datasets

Loss	ALOI			GRAZ02			NUS-WIDE-v1			NUS-WIDE-v2		
	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
$\mathscr{L}_{Re}$	71.36 –	82.16 –	76.09 –	38.28 –	7.30 –	54.47 –	27.00 –	12.05 –	35.06 –	14.40 –	18.29 –	25.85 –
$\mathscr{L}_{Re}+\mathscr{L}_{Sub}$	78.68 ↑	81.16 ↓	72.47 ↓	39.77 ↑	8.74 ↑	55.08 ↑	30.69 ↑	19.82 ↑	36.18 ↑	15.35 ↑	16.58 ↓	27.40 ↑
$\mathscr{L}_{Re}+\mathscr{L}_{Sub}+\mathscr{L}_{Con}$	88.32 ↑	84.17 ↑	79.98 ↑	45.33 ↑	12.46 ↑	56.23 ↑	32.50 ↑	19.88 ↑	36.88 ↑	15.80 ↑	17.59 ↓	25.75 ↓
$\mathscr{L}_{Re}+\mathscr{L}_{Sub}+\mathscr{L}_{Fu}$	93.05 ↑	89.87 ↑	84.52 ↑	45.12 ↑	11.18 ↑	57.18 ↑	32.88 ↑	18.82 ↑	36.88 ↑	15.70 ↑	19.03 ↑	27.80 ↑
$\mathscr{L}_{Sub}+\mathscr{L}_{Con}+\mathscr{L}_{Fu}$	89.62 ↑	86.58 ↑	85.82 ↑	45.39 ↑	11.66 ↑	58.27 ↑	33.00 ↑	17.28 ↑	36.12 ↑	15.45 ↑	18.92 ↑	26.90 ↑
$\mathscr{L}$	**95.74** ↑	**92.72** ↑	**95.74** ↑	**51.90** ↑	**16.16** ↑	**59.55** ↑	**36.56** ↑	**21.83** ↑	**40.06** ↑	**17.85** ↑	**21.23** ↑	**30.30** ↑

The best results are bolded.

Table 13. Ablation Results of the Proposed SCMC on Four Datasets

Loss	Reuters			UCI			WikipediaArticles			Youtube		
	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
$\mathscr{L}_{Re}$	41.80 –	17.53 –	45.93 –	70.90 –	69.16 –	73.90 –	23.09 –	9.59 –	29.73 –	23.95 –	13.00 –	29.15 –
$\mathscr{L}_{Re}+\mathscr{L}_{Sub}$	38.33 ↓	28.52 ↑	51.07 ↑	75.65 ↑	73.88 ↑	78.65 ↑	35.64 ↑	24.54 ↑	40.69 ↑	27.65 ↑	17.65 ↑	32.40 ↑
$\mathscr{L}_{Re}+\mathscr{L}_{Sub}+\mathscr{L}_{Con}$	47.60 ↑	31.05 ↑	49.93 ↑	91.40 ↑	83.70 ↑	81.90 ↑	37.95 ↑	27.72 ↑	41.99 ↑	29.20 ↑	18.74 ↑	35.40 ↑
$\mathscr{L}_{Re}+\mathscr{L}_{Sub}+\mathscr{L}_{Fu}$	48.00 ↑	29.34 ↑	51.00 ↑	95.55 ↑	90.99 ↑	86.25 ↑	56.57 ↑	**54.67** ↑	62.36 ↑	31.90 ↑	18.90 ↑	37.60 ↑
$\mathscr{L}_{Sub}+\mathscr{L}_{Con}+\mathscr{L}_{Fu}$	48.60 ↑	30.10 ↑	57.53 ↑	93.00 ↑	86.78 ↑	93.15 ↑	52.23 ↑	51.89 ↑	62.04 ↑	34.40 ↑	21.17 ↑	38.75 ↑
$\mathscr{L}$	**51.80** ↑	**34.47** ↑	**58.13** ↑	**96.75** ↑	**92.84** ↑	**97.00** ↑	**57.86** ↑	53.74 ↑	**63.20** ↑	**37.90** ↑	**26.12** ↑	**41.15** ↑

The best results are bolded.

Table 14. Ablation Results of the Proposed SCMC on Two Datasets

Loss	Animals			Cifar10		
	ACC	NMI	Purity	ACC	NMI	Purity
$\mathscr{L}_{Re}$	48.10 –	63.59 –	53.94 –	15.73 –	5.28 –	22.55 –
$\mathscr{L}_{Re} + \mathscr{L}_{Sub}$	50.34 ↑	65.87 ↑	58.46 ↑	17.10 ↑	5.60 ↑	23.63 ↑
$\mathscr{L}_{Re} + \mathscr{L}_{Sub} + \mathscr{L}_{Con}$	54.84 ↑	66.60 ↑	63.20 ↑	18.26 ↑	7.00 ↑	23.57 ↑
$\mathscr{L}_{Re} + \mathscr{L}_{Sub} + \mathscr{L}_{Fu}$	54.50 ↑	66.96 ↑	62.47 ↑	18.36 ↑	8.08 ↑	24.78 ↑
$\mathscr{L}_{Sub} + \mathscr{L}_{Con} + \mathscr{L}_{Fu}$	52.72 ↑	68.02 ↑	62.90 ↑	18.40 ↑	7.34 ↑	24.90 ↑
$\mathscr{L}$	**57.06** ↑	**68.82** ↑	**64.16** ↑	**21.22** ↑	**10.26** ↑	**25.56** ↑

The best results are bolded.

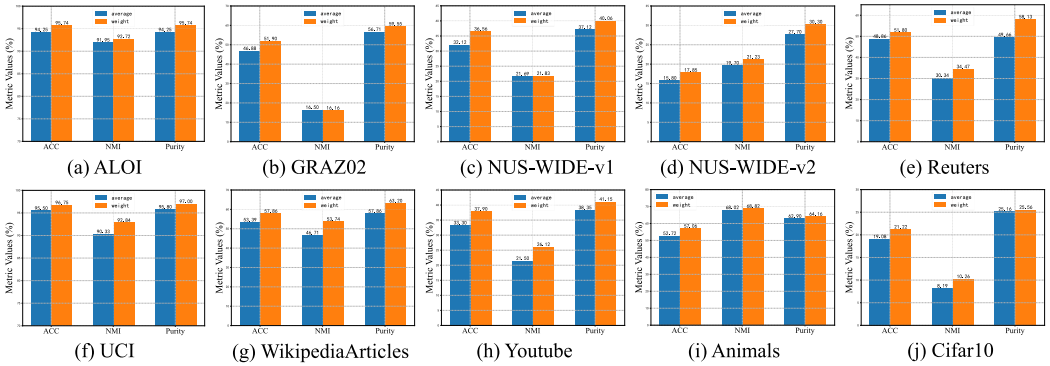


Fig. 10. The comparison of clustering results with weight fusion and average fusion.

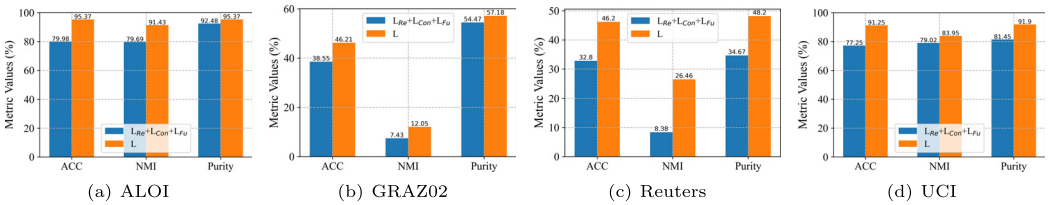
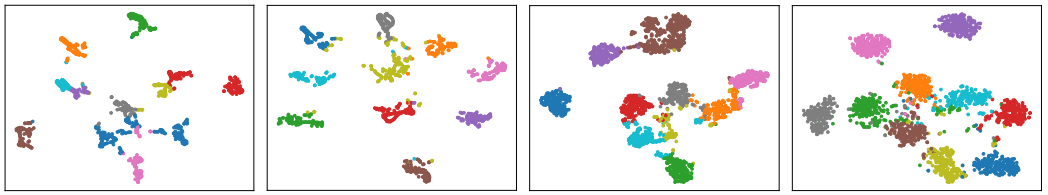


Fig. 11. The comparison of clustering results with respect to latent space contrast and subspace contrast.

$$\begin{aligned}\mathscr{L}_{Fu} &= \min_{\mathbf{C}^{(v)},\mathbf{A}} \sum_{v}\sum_{i}\sum_{j} ||\mathbf{C}_i^{(v)} - \mathbf{C}_j^{(v)}||_2^2 \mathbf{A}_{ij} + \sum_{i}\sum_{j} \mathbf{A}_{ij}^2 \\ &= \sum_{v=1}^{V} Tr(\mathbf{C}^{(v)} \mathbf{L}_{\mathbf{A}} \mathbf{C}^{(v)^T}) + ||\mathbf{A}||_F^2 \\ &\text{s.t. } \mathbf{A}_i \mathbf{1} = 1, \mathbf{A}_{ij} \geq 0, \mathbf{A}_{ii} = 0,\end{aligned} \tag{34}$$

where $\mathbf{C}^{(v)}$ is the latent representation learned via $\mathbf{C}^{(v)} = e_v(\mathbf{X}^{(v)}|\mathbf{W}_e^{(v)}, \mathbf{b}_e^{(v)})$. Besides, the weighted fusion scheme is discarded. The designed ablation method is tentatively named **Latent-Contrastive Multi-view Clustering (LCMC)**. After the training ends, the clustering results are obtained via performing the spectral clustering algorithm on the affinity matrix $(\mathbf{A} + \mathbf{A}^T)/2$. Certainly, to maintain the fairness of experiments, we also train the proposed SCMC without the weighted fusion mechanism. Figure 11 shows the result comparisons under the two contrast methods. Obviously, when executing the contrast at the subspace level, the experimental performance is much better that when executing the contrast between latent representations.

Secondly, to further avoid the influence of fusion loss $\mathscr{L}_{Fu}$, we remove the fusion loss $\mathscr{L}_{Fu}$ of LCMC and SCMC, and obtain the unified latent and subspace representation via $\mathbf{C} = \sum_{v=1}^{V} \mathbf{C}^{(v)}/V$, $\mathbf{Z} = \sum_{v=1}^{V} \mathbf{Z}^{(v)}/V$, respectively. Thus, the clustering results are acquired by adopting the spectral clustering algorithm. We leverage the t-SNE method to reduce the dimensionality of $\mathbf{C}$ and $\mathbf{Z}$ learned on ALOI and UCI datasets, respectively, and visualize the scatter plots. As Figure 12 shows, many different clusters visualized by latent features adhesive with each other, the overall separation is not good. On the contrary, the cluster shapes visualized by the subspace features are distinct, and diverse clusters are pulled apart. This phenomenon illustrates the subspace features are more



(a) ALOI: latent features (b) ALOI: subspace features (c) UCI: latent features (d) UCI: subspace features

Fig. 12. Scatter comparison using latent features and subspace features on ALOI and UCI datasets via the t-SNE technology.

Table 15. Clustering Performance of the Compared Methods on ALOI Dataset with In-Samples and Out-of-Samples

Method	In-Samples						Out-of-Samples					
	ACC	NMI	Purity	ARI	F-score	Precision	ACC	NMI	Purity	ARI	F-score	Precision
K-means	47.63	48.39	48.63	34.28	42.22	34.66	49.10	49.02	52.69	30.45	38.50	33.06
SwMC	67.88	65.28	71.00	38.32	46.45	34.95	70.61	65.52	70.61	40.80	48.86	35.65
MCGC	80.38	72.97	80.37	63.38	67.16	64.62	81.00	77.95	81.00	68.86	72.24	66.19
LMVSC	71.00	73.21	71.75	63.31	67.28	61.69	73.12	73.56	74.91	58.57	63.34	54.67
SCMC	**96.13**	**93.72**	**96.12**	**91.55**	**93.10**	**93.25**	**93.55**	**90.53**	**93.55**	**87.47**	**88.72**	**88.37**

The best results are bolded.

Table 16. Clustering Performance of the Compared Methods on GRAZ02 Dataset with In-Samples and Out-of-Samples

Method	In-Samples						Out-of-Samples					
	ACC	NMI	Purity	ARI	F-score	Precision	ACC	NMI	Purity	ARI	F-score	Precision
K-means	36.56	3.40	36.56	3.93	33.55	27.42	36.67	4.69	37.33	3.57	33.36	27.18
SwMC	42.86	11.76	46.34	8.93	36.29	30.44	41.00	7.63	42.67	5.36	33.32	28.39
MCGC	45.15	9.16	45.15	8.75	31.74	31.75	41.67	8.35	44.00	7.79	30.94	31.01
LMVSC	42.35	8.01	42.35	7.54	31.24	30.71	42.33	6.09	42.33	5.90	30.65	29.33
SCMC	**49.82**	**15.03**	**59.77**	**13.73**	**36.43**	**36.90**	**49.66**	**15.55**	**59.66**	**13.95**	**36.75**	**37.19**

The best results are bolded.

discriminative than the latent features, which means that the subspace learning does alleviate the modality separation and help the contrastive training.

5.8 Experiments on Out-of-Samples

To verify the effectiveness of the proposed CSR, we plug it into several MVC methods SwMC, CSMSC, LMVSC, and our SCMC to conduct the experiments. Specifically, 80% of all samples in a multi-view dataset are randomly selected to consist of the in-samples, and the remaining 20% consist of the out-of-samples. *k*-means is used as the most basic compared method. Actually, *k*-means does not have the ability to handle the out-of-samples, so we have to perform it on the in-samples and out-of-samples and obtain the clustering results, respectively. As for the other compared methods, we first perform them on the in-samples to obtain their labels, then adopt the proposed CSR mechanism on the out-of-samples to generate their labels. Tables 15–18 report the experimental results of several compared methods on four multi-view datasets. It can be seen that the proposed

Table 17. Clustering Performance of the Compared Methods on Reuters Dataset with In-Samples and Out-of-Samples

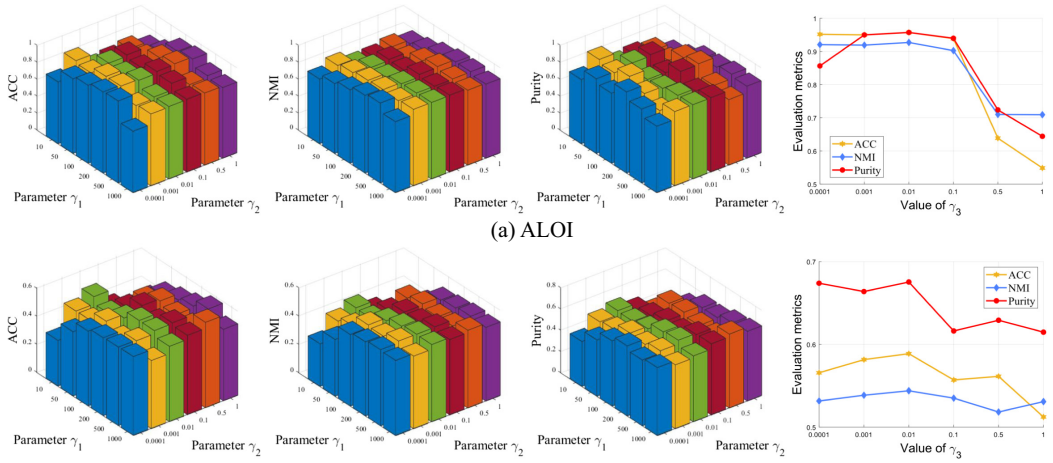
Method	In-Samples						Out-of-Samples					
	ACC	NMI	Purity	ARI	F-score	Precision	ACC	NMI	Purity	ARI	F-score	Precision
K-means	36.75	4.26	36.92	5.47	35.22	24.08	41.67	9.11	42.67	10.20	37.91	26.94
SwMC	47.75	28.40	53.75	13.48	38.97	28.46	45.00	26.62	51.67	5.96	34.64	24.90
MCGC	46.33	18.01	51.50	18.66	40.31	32.58	44.67	22.92	53.67	15.56	37.65	31.56
LMVSC	46.00	24.40	50.50	18.93	41.19	32.28	47.33	27.23	52.33	16.02	38.47	31.55
SCMC	**50.25**	**29.24**	**56.75**	**21.38**	**49.2**	**41.37**	**51.00**	**30.88**	**58.33**	**18.21**	**41.88**	**31.73**

The best results are bolded.

Table 18. Clustering Performance of the Compared Methods on UCI Dataset with In-Samples and Out-of-Samples

Method	In-Samples						Out-of-Samples					
	ACC	NMI	Purity	ARI	F-score	Precision	ACC	NMI	Purity	ARI	F-score	Precision
K-means	39.12	47.01	44.56	31.93	39.41	35.75	37.75	45.03	41.25	26.25	34.52	30.88
SwMC	73.25	79.49	75.94	68.05	71.58	64.58	75.25	78.99	77.50	66.10	70.06	60.66
MCGC	83.06	78.77	83.06	72.63	75.37	74.86	82.50	78.62	83.00	72.75	75.65	71.43
LMVSC	89.56	82.51	89.56	79.35	81.42	80.96	87.50	82.44	87.50	76.12	78.54	77.65
SCMC	**95.56**	**90.75**	**96.19**	**90.41**	**91.49**	**91.51**	**95.00**	**90.56**	**95.00**	**88.82**	**89.93**	**90.15**

The best results are bolded.



(a) ALOI

(b) WikipediaArticles

Fig. 13. The parameter sensitivity analysis of the proposed SCMC on ALOI and WikipediaArticles datasets.

SCMC achieves the optimal performance on both in-samples and out-of-samples. Moreover, the proposed CSR has excellent adaptability and can be combined with any MVC algorithms to yield the superior performance on the out-of-samples.

5.9 Parameter Sensitivity Analysis

Three nonnegative tradeoff parameters are used to balance the overall objective loss, their impacts on the clustering outcomes are investigated via sensitivity experiments. Specifically, we tune γ_1, γ_2,

and γ_3 in {10, 50, 100, 200, 500, 1, 000}, {0.0001, 0.001, 0.01, 0.1, 0.5, 1}, and {0.0001, 0.001, 0.01, 0.1, 0.5, 1}, respectively. Figure 13 shows the numerical results of SCMC under different settings of γ_1, γ_2, and γ_3. When evaluating γ_1 and γ_2, we fix γ_3, the same operation is done for evaluating γ_3. It can been seen that a small γ_1 is not conducive to learning an informative subspace representation, this is caused by the insufficient penalty for $||\mathbf{C}^{(v)^T} - \mathbf{C}^{(v)^T}\mathbf{Z}^{(v)}||_F^2$ in the optimization. Similarly, when γ_2 is set to be small, the clustering performance is not ideal, because the contrast mechanism has little effects to enhance the discrimination of subspace representations. Interestingly, selecting a relatively lager γ_3 can degrade the effectiveness of SCMC, this situation may result from over-smoothing.

6 Conclusion

In this paper, we propose a novel SCMC approach. In SCMC, the subspace representation of each view is nonlinearly explored through a extraction network. Inspired by the idea of contrastive learning, we regard the subspace representation of each view as a contrastable entity. By pairwise comparison of multiple subspace representations, we exploit the complementary information in them and augment the discriminability of each subspace representation. Guided by the important principle of multi-view consensus, we obtain a consistent affinity matrix by the graph regularization. Furthermore, to handle the out-of-samples, the CSR learning method over the in-samples is proposed. In future work, we focus on three important challenges in deep multi-view subspace clustering. First, how to achieve more effective contrast between subspace representations. The proposed SCMC uses a cross-view comparison approach while ignoring the intra-view comparison relationship, which needs to be paid attention to. Second, how to improve the efficiency of subspace representation learning and reduce its memory overhead. Self-expression based subspace representation learning suffers from huge time and space overhead, which severely limits its practical applications, and the development of efficient multi-view subspace learning methods is necessary. Third, how to deal with possible missing values in multi-view data. Due to various negative problems such as sensor failures, multi-view data are often incomplete, which affects the learning of subspace representations, it is significant to study a multi-view subspace learning algorithm that can adaptively deal with missing data.

Appendices

A The Definition of Evaluation Metrics

Following the compared methods [28, 31, 36, 55, 65], we select six frequently-used clustering evaluation metrics to evaluate the clustering results, including ACC, NMI, Purity, ARI, F-score, and Presion. To elaborate on which aspect of the clustering results they characterize respectively, their definitions need to be introduced.

Suppose y_i is the ith sample's true label, $\hat{y}_i$ is the ith sample's predicted label, and $\Phi(\cdot)$ is the best mapping function from the predicted label to the true label, then ACC is defined as

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \psi\left(y_i, \Phi\left(\hat{y}_i\right)\right)}{N}, \tag{35}$$

where N is the number of instances, $\psi(\cdot,\cdot)$ is the binary criterion function and defined as

$$\psi(x, y) = \begin{cases} 1, & \text{if } x = y; \\ 0, & \text{otherwise.} \end{cases}$$

Given that $Y = \{Y_i\}_{i=1}^{c}$ is the true clusters and $\hat{Y} = \{\hat{Y}_j\}_{j=1}^{\hat{c}}$ is the predicted clusters, then NMI is defined as

$$\text{NMI} = \frac{I(\hat{Y}, Y)}{(F(\hat{Y}) + F(Y))/2}, \tag{36}$$

where $I(\cdot,\cdot)$ denotes the mutual information between two variables, $F(\cdot)$ denotes the entropy of certain variable. The mutual information $I(\cdot,\cdot)$ is computed by

$$I(\hat{Y}, Y) = \sum_{j=1}^{\hat{c}} \sum_{i=1}^{c} P(\hat{Y}_i \cap Y_j) log \frac{P(\hat{Y}_i \cap Y_j)}{P(\hat{Y}_i)P(Y_j)}, \tag{37}$$

where $P(\cdot)$ denotes the probability. The entropy $F(\hat{Y})$, $F(Y)$ are respectively computed by

$$F(\hat{Y}) = -\sum_{j=1}^{\hat{c}} P(\hat{Y}_j) log(P(\hat{Y}_j)), F(Y) = -\sum_{i=1}^{c} P(Y_i) log(P(Y_i)). \tag{38}$$

Purity indicates the ratio of correctly grouped samples and is calculated as

$$\text{Purity} = \frac{1}{N} \sum_{j=1}^{\hat{c}} max_i |\hat{Y}_j \cap Y_i| \tag{39}$$

Assume that TP, FP, TN, and FN represent the number of entries correctly predicted as positive instances, the number of entries erroneously predicted as positive instances, the number of entries correctly predicted as negative instances, the number of erroneously predicted as negative instances, respectively. Thus, the definition of **Rand Index (RI)** is written as

$$\text{RI} = \frac{TP + TN}{TP + FP + TF + FN}. \tag{40}$$

RI cannot guarantee that the evaluation value of clustering results of randomized division tends to be zero, ARI is introduced as follows,

$$\begin{aligned} \text{ARI} &= \frac{\text{RI} - E[\text{RI}]}{max(\text{RI}) - EI[\text{RI}]} \\ &= \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}, \end{aligned} \tag{41}$$

where n_{ij} denotes the number of samples overlap between the true ith cluster and the predicted jth cluster, a_i denotes the number of samples belonging to the ith true cluster, b_j denotes the number of samples belonging to the jth predicted cluster.

Precision denotes the proportion of correctly predicted positive samples to all predicted positive samples, which is defined as

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{42}$$

Recall denotes the proportion of correctly predicted positive samples to all true positive samples, which is defined as

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{43}$$

To balance the Precision and Recall, the F-score is proposed and defined as

$$\text{F-score} = 2\frac{Precision * Recall}{Precision + Recall}. \tag{44}$$

Table 19. Characteristics of Six Metrics

Metric	Characteristic
ACC	Computes the clustering accuracy from a overall view, but is susceptible to the class imbalance problem.
NMI	Measures the similarity degree between the predicted labels and true labels.
Purity	Computes the ratio of correctly clustered samples, but cannot balance the quality of clustering against the number of clusters.
ARI	Measures the similarity degree between the predicted labels and true labels.
Precision	Computes the proportion of correctly predicted positive samples to all predicted positive samples for each category, then averages the Precision values of all categories.
F-score	Computes proportion of correctly predicted positive samples, is a balanced solution of Precision and Recall, and adapts to the situation of class imbalance.

Table 20. Clustering Results on Each View of Four Datasets Via Spectral Clustering

View Index (Method)	ALOI			GRAZ02			NUS-WIDE-v1			NUS-WIDE-v2		
	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
View 1 (SC)	57.58	71.98	67.82	35.37	3.80	37.26	13.19	0.85	13.38	12.30	1.83	13.85
View 2 (SC)	62.09	62.05	64.23	36.04	9.11	40.11	19.38	4.03	20.19	12.35	1.53	13.25
View 3 (SC)	58.33	59.59	60.50	28.52	0.23	28.66	17.69	3.65	19.38	12.65	1.85	13.70
View 4 (SC)	51.96	51.04	54.46	28.52	0.23	28.66	15.69	2.62	16.81	12.50	1.82	13.40
View 5 (SC)	–	–	–	35.77	9.08	20.94	20.94	8.86	24.75	13.75	4.17	14.80
View 6 (SC)	–	–	–	36.11	5.88	38.21	14.75	7.75	19.75	–	–	–
– (SCMC)	**95.74**	**92.12**	**95.74**	**51.90**	**16.16**	**59.55**	**36.56**	**21.83**	**40.06**	**17.85**	**21.23**	**30.30**

The best results are bolded.

Table 21. Clustering Results on Each View of Four Datasets Via Spectral Clustering

View Index (Method)	Reuters			UCI			WikipediaArticles			Youtube		
	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
View 1 (SC)	28.00	0.39	28.13	30.55	33.42	31.90	19.77	6.55	21.93	22.20	8.74	22.65
View 2 (SC)	27.93	0.34	28.07	70.95	64.10	70.95	55.12	51.91	60.03	22.40	10.51	24.60
View 3 (SC)	27.93	0.34	28.07	22.25	11.46	23.90	–	–	–	19.20	8.38	20.70
View 4 (SC)	26.20	0.38	28.60	–	–	–	–	–	–	30.75	20.22	33.15
View 5 (SC)	27.80	0.28	28.00	–	–	–	–	–	–	29.90	16.91	31.95
View 6 (SC)	–	–	–	–	–	–	–	–	–	27.55	15.16	29.80
– (SCMC)	**51.80**	**34.47**	**58.13**	**96.75**	**92.84**	**97.00**	**57.86**	**53.74**	**63.20**	**37.90**	**26.12**	**41.15**

The best results are bolded.

It is worthy emphasizing that when calculating the overall values of Precision, Recall, and F-score, the values for each class are first calculated separately and then averaged. Furthermore, Recall focuses on the retrieval rate of correctly categorized samples. Metaphorically speaking, it is preferred to misclassify more samples than to miss one. In the scenario of multi-class clustering, the Recall metric is easily biased by certain categories and cannot objectively reflect the sample division, so we discard the metric.

Ultimately, we summarize the characteristics of the six metrics in the following Table 19.

B The Verification of Capturing the Complementary Information

To demonstrate that the proposed SCMC can capture the complementary information in different views, we execute the spectral clustering algorithm on each view of the test multi-view datasets and report the clustering results in Tables 20–22. It can be observed that the clustering results under any of the views are worse than those obtained by the proposed SCMC, which is a good indication that SCMC captures the complementarity information among multiple views.

Table 22. Clustering Results on Each View of Two Datasets Via Spectral Clustering

View Index (Method)	Animals			Cifar10		
	ACC	NMI	Purity	ACC	NMI	Purity
View 1 (SC)	28.88	36.63	33.18	10.46	0.10	10.54
View 2 (SC)	33.76	51.51	40.98	11.92	4.61	15.34
– (SCMC)	**57.06**	**68.82**	**64.16**	**21.22**	**10.26**	**25.56**

The best results are bolded.

References

[1] Arthur Asuncion and David Newman. 2007. UCI machine learning repository.

[2] Yoshua Bengio, Jean-François Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. 2003. Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering. In *Proceedings of the Advances in Neural Information Processing Systems*. 177–184.

[3] Jinyu Cai, Jicong Fan, Wenzhong Guo, Shiping Wang, Yunhe Zhang, and Zhao Zhang. 2022. Efficient deep embedded subspace clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1–10.

[4] Xiaosha Cai, Dong Huang, Guang-Yu Zhang, and Chang-Dong Wang. 2023. Seeking commonness and inconsistencies: A jointly smoothed approach to multi-view subspace clustering. *Information Fusion* 91 (2023), 364–375.

[5] Guoqing Chao, Shiliang Sun, and Jinbo Bi. 2021. A survey on multiview clustering. *IEEE Transactions on Artificial Intelligence* 2, 2 (2021), 146–168.

[6] Guoqing Chao, Songtao Wang, Shiming Yang, Chunshan Li, and Dianhui Chu. 2022. Incomplete multi-view clustering with multiple imputation and ensemble clustering. *Applied Intelligence* 52, 13 (2022), 14811–14821.

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

[8] Zhaoliang Chen, Lele Fu, Shunxin Xiao, Shiping Wang, Claudia Plant, and Wenzhong Guo. 2023. Multi-view graph convolutional networks with differentiable node selection. *ACM Transactions on Knowledge Discovery from Data* 18 (2023), 1–21.

[9] Lei Cheng, Yongyong Chen, and Zhongyun Hua. 2022. Deep Contrastive Multi-view Subspace Clustering. In *Proceedings of the International Conference on Neural Information Processing*. 692–704.

[10] Beilei Cui, Hong Yu, Linlin Zong, and Ziyang Cheng. 2021. Self-guided deep multi-view subspace clustering network. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. 1–6.

[11] Guowang Du, Lihua Zhou, Kevin Lü, Hao Wu, and Zhimin Xu. 2023. Multiview subspace clustering with multilevel representations and adversarial regularization. *IEEE Transactions on Neural Networks and Learning Systems* 34, 12 (2023), 10279–10293. DOI: https://doi.org/10.1109/TNNLS.2022.3165542

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

[13] Ehsan Elhamifar and Rene Vidal. 2013. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 11 (2013), 2765–2781.

[14] Lele Fu, Zhaoliang Chen, Yongyong Chen, and Shiping Wang. 2023. Unified low-rank tensor learning and spectral embedding for multi-view subspace clustering. *IEEE Transactions on Multimedia* 25 (2023), 4972–4985.

[15] Lele Fu, Pengfei Lin, Athanasios V. Vasilakos, and Shiping Wang. 2020. An overview of recent multi-view clustering. *Neurocomputing* 402 (2020), 148–161.

[16] Hongchang Gao, Feiping Nie, Xuelong Li, and Heng Huang. 2015. Multi-view subspace clustering. In *Proceedings of the IEEE International Conference on Computer Vision*. 4238–4246.

[17] Jing Gao, Meng Liu, Peng Li, Jianing Zhang, and Zhikui Chen. 2023. Deep multiview adaptive clustering with semantic invariance. *IEEE Transactions on Neural Networks and Learning Systems* (2023). DOI: 10.1109/TNNLS.2023.3265699.

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *Proceedings of the Advances in Neural Information Processing Systems* 33 (2020), 21271–21284.

[19] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. 2017. Improved deep embedded clustering with local structure preservation. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*. 1753–1759.

[20] Kaveh Hassani and Amir H. Khasahmadi. 2020. Contrastive multi-view representation learning on graphs. In *Proceedings of the International Conference on Machine Learning*. 4116–4126.
[21] Zhanxuan Hu, Feiping Nie, Rong Wang, and Xuelong Li. 2020. Multi-view spectral clustering via integrating nonnegative embedding and spectral embedding. *Information Fusion* 55 (2020), 251–259.
[22] Sheng Huang, Yunhe Zhang, Lele Fu, and Shiping Wang. 2022. Learnable multi-view matrix factorization with graph embedding and flexible loss. *IEEE Transactions on Multimedia* 25 (2022), 3259–3272. DOI: 10.1109/TMM.2022.3157997.
[23] Jintian Ji and Songhe Feng. 2023. Anchor structure regularization induced multi-view subspace clustering via enhanced tensor rank minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19343–19352.
[24] Zhao Kang, Wangtao Zhou, Zhitong Zhao, Junming Shao, Meng Han, and Zenglin Xu. 2020. Large-scale multi-view subspace clustering in linear time. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*. 4412–4419.
[25] Ao Li, Cong Feng, Yuan Cheng, Yingtao Zhang, and Hailu Yang. 2023. Incomplete multiview subspace clustering based on multiple kernel low-redundant representation learning. *Information Fusion* 103 (2023), 102086.
[26] Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. 2023. Incomplete multi-view clustering via prototype-based imputation. arXiv:2301.11045. Retrieved from https://doi.org/10.48550/arXiv.2301.11045
[27] Yunfan Li, Peng Hu, Jerry Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8547–8555.
[28] Zhenglai Li, Chang Tang, Xinwang Liu, Xiao Zheng, Guanghui Yue, Wei Zhang, and En Zhu. 2022. Consensus graph learning for multi-view clustering. *IEEE Transactions on Multimedia* 24 (2022), 2461–2472.
[29] Zhenglai Li, Chang Tang, Xiao Zheng, Xinwang Liu, Wei Zhang, and En Zhu. 2022. High-order correlation preserved incomplete multi-view subspace clustering. *IEEE Transactions on Image Processing* 31 (2022), 2067–2080.
[30] Hongfu Liu and Yun Fu. 2018. Consensus guided multi-view clustering. *ACM Transactions on Knowledge Discovery from Data* 12, 4 (2018), 4201–4221.
[31] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Xifeng Guo, Marius Kloft, and Liangzhong He. 2021. Multiview subspace clustering via co-training robust data representation. *IEEE Transactions on Neural Networks and Learning Systems* 33, 10 (2021), 5177–5189.
[32] Suyuan Liu, Siwei Wang, Pei Zhang, Xinwang Liu, Kai Xu, Changwang Zhang, and Feng Gao. 2022. Efficient one-pass multi-view subspace clustering with consensus anchors. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 7576–7584
[33] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. 2017. A survey of deep neural network architectures and their applications. *Neurocomputing* 234 (2017), 11–26.
[34] Yue Liu, Xihong Yang, Sihang Zhou, Xinwang Liu, Siwei Wang, Ke Liang, Wenxuan Tu, and Liang Li. 2023. Simple contrastive graph clustering. *IEEE Transactions on Neural Networks and Learning Systems* (2023), 1–12. DOI: https://doi.org/10.1109/TNNLS.2023.3271871
[35] Zhoumin Lu, Feiping Nie, Rong Wang, and Xuelong Li. 2022. A differentiable perspective for multi-view spectral clustering with flexible extension. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2022), 7087–7098.
[36] Shirui Luo, Changqing Zhang, Wei Zhang, and Xiaochun Cao. 2018. Consistent and specific multi-view subspace clustering. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*. 3730–3737.
[37] Feiping Nie, Jing Li, and Xuelong Li. 2016. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*. 1881–1887.
[38] Feiping Nie, Jing Li, and Xuelong Li. 2017. Self-weighted multiview clustering with multiple graphs. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*. 2564–2570.
[39] Feiping Nie, Xiaoqian Wang, and Heng Huang. 2014. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. 977–986.
[40] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. 2019. COMIC: Multi-view clustering without parameter selection. In *Proceedings of the International conference on machine learning*. 5092–5101.
[41] Xi Peng, Lei Zhang, and Zhang Yi. 2013. Scalable sparse subspace clustering. In *Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition*. 430–437.
[42] Yalan Qin, Nan Pu, and Hanzhou Wu. 2023. Elastic Multi-view Subspace Clustering with Pairwise and High-order Correlations. *IEEE Transactions on Knowledge and Data Engineering* (2023), 1–13. DOI: 10.1109/TKDE.2023.3293498
[43] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. 2023. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5571–5584.
[44] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61 (2015), 85–117.

[45] Xiukun Sun, Miaomiao Cheng, Chen Min, and Liping Jing. 2019. Self-supervised deep multi-view subspace clustering. In *Proceedings of the Asian Conference on Machine Learning*. 1001–1016.
[46] Yuze Tan, Yixi Liu, Shudong Huang, Wentao Feng, and Jiancheng Lv. 2023. Sample-level multi-view graph clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23966–23975.
[47] Chang Tang, Zhenglai Li, Jun Wang, Xinwang Liu, Wei Zhang, and En Zhu. 2023. Unified one-step multi-view spectral clustering. *IEEE Transactions on Knowledge and Data Engineering* 35, 6 (2023), 6449–6460. DOI: 10.1109/TKDE.2022.3172687
[48] Daniel J. Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. 2021. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1255–1265.
[49] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11 (2010), 3371–3408.
[50] Qianqian Wang, Jiafeng Cheng, Quanxue Gao, Guoshuai Zhao, and Licheng Jiao. 2021. Deep multi-view subspace clustering with unified and discriminative learning. *IEEE Transactions on Multimedia* 23 (2021), 3483–3493.
[51] Shuqin Wang, Yongyong Chen, Zhiping Lin, Yigang Cen, and Qi Cao. 2023. Robustness meets low-rankness: Unified entropy and tensor learning for multi-view subspace clustering. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 11 (2023), 6302–6316.
[52] Shuqin Wang, Yongyong Chen, Shuang Yi, and Guoqing Chao. 2022. Frobenius norm-regularized robust graph learning for multi-view subspace clustering. *Applied Intelligence* 52, 13 (2022), 14935–14948.
[53] Shiye Wang, Changsheng Li, Yanming Li, Ye Yuan, and Guoren Wang. 2023. Self-supervised information bottleneck for deep multi-view subspace clustering. *IEEE Transactions on Image Processing* 32 (2023), 1555–1567.
[54] Yu Wang, Chuan Chen, Jinrong Lai, Lele Fu, Yuren Zhou, and Zibin Zheng. 2023. A self-representation method with local similarity preserving for fast multi-view outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 17, 1 (2023), 1–20. DOI: 10.1145/3532191
[55] Wei Xia, Quanxue Gao, Qianqian Wang, Xinbo Gao, Chris Ding, and Dacheng Tao. 2022. Tensorized bipartite graph learning for multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 5187–5202.
[56] Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the International Conference on Machine Learning*, Vol. 48. 478–487.
[57] Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. arXiv:1304.5634.
[58] Jie Xu, Huayi Tang, Yazhou Ren, Xiaofeng Zhu, and Lifang He. 2022. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 16051–16060.
[59] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. 2006. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 1 (2006), 40–51.
[60] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. 2023. GCFAgg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19863–19872.
[61] Ben Yang, Xuetao Zhang, Feiping Nie, Fei Wang, Weizhong Yu, and Rong Wang. 2021. Fast multi-view clustering via nonnegative and orthogonal factorization. *IEEE Transactions on Image Processing* 30 (2021), 2575–2586.
[62] J. H. Yang, C. Chen, H. N. Dai, M. Ding, Z. B. Wu, and Z. B. Zheng. 2022. Robust corrupted data recovery and clustering via generalized transformed tensor low-rank representation. *IEEE Trans. Neural Networks Learning Systems* (2022). DOI: 10.1109/TNNLS.2022.3215983
[63] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. 2022. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 1055–1069.
[64] Z. Yang, Q. Xu, W. Zhang, X. Cao, and Q. Huang. 2019. Split multiplicative multi-view subspace clustering. *IEEE Transactions on Image Processing* 28, 10 (2019), 5147–5160.
[65] Kun Zhan, Feiping Nie, Jing Wang, and Li Yang. 2019. Multiview consensus graph clustering. *IEEE Transactions on Image Processing* 28, 3 (2019), 1261–1270.
[66] Lei Zhang, Lele Fu, Tong Wang, Chuan Chen, and Chuanfu Zhang. 2023. Mutual information-driven multi-view clustering. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 3268–3277.
[67] Xu Zhang, Wen Wang, Zhe Chen, Yufei Xu, Jing Zhang, and Dacheng Tao. 2023. CLAMP: Prompt-based contrastive learning for connecting language and animal pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23272–23281.
[68] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. 2017. Multi-view learning overview: Recent progress and new challenges. *Information Fusion* 38 (2017), 43–54.

[69] Pengfei Zhu, Binyuan Hui, Changqing Zhang, Dawei Du, Longyin Wen, and Qinghua Hu. 2019. Multi-view deep subspace clustering networks. arXiv:1908.01978. Retrieved from https://doi.org/10.48550/arXiv.1908.01978

[70] Xiaofeng Zhu, Shichao Zhang, Yonghua Zhu, Wei Zheng, and Yang Yang. 2020. Self-weighted multi-view fuzzy clustering. *ACM Transactions on Knowledge Discovery from Data* 14 (2020), 4801–4817.

Received 6 March 2023; revised 1 April 2024; accepted 17 June 2024