

Deep Multiview Clustering by Contrasting Cluster Assignments

Jie Chen¹, Hua Mao², Wai Lok Woo², Xi Peng^{1*}

¹ College of Computer Science, Sichuan University, China

² Department of Computer and Information Sciences, Northumbria University

chenjie2010@scu.edu.cn; {hua.mao, wailok.woo}@northumbria.ac.uk;

pengx.gm@gmail.com

Abstract

Multiview clustering (MVC) aims to reveal the underlying structure of multiview data by categorizing data samples into clusters. Deep learning-based methods exhibit strong feature learning capabilities on large-scale datasets. For most existing deep MVC methods, exploring the invariant representations of multiple views is still an intractable problem. In this paper, we propose a cross-view contrastive learning (CVCL) method that learns view-invariant representations and produces clustering results by contrasting the cluster assignments among multiple views. Specifically, we first employ deep autoencoders to extract view-dependent features in the pretraining stage. Then, a cluster-level CVCL strategy is presented to explore consistent semantic label information among the multiple views in the fine-tuning stage. Thus, the proposed CVCL method is able to produce more discriminative cluster assignments by virtue of this learning strategy. Moreover, we provide a theoretical analysis of soft cluster assignment alignment. The extensive experimental results obtained on several datasets demonstrate that the proposed CVCL method outperforms several state-of-the-art approaches.

1. Introduction

Multiview data are usually represented by different types of features or collected from multiple sources. All views share the same semantic information contained in the multiview data. Simultaneously, the data information derived from multiple views is complementary [6, 36]. The goal of multiview clustering (MVC) is to divide data samples into different groups according to their distinct feature information.

MVC has attracted increasing attention for many machine learning tasks, including feature selection [36], scene recognition [23] and information retrieval [8, 41, 5]. The ex-

isting literature involving traditional machine learning techniques can be roughly divided into four categories, including subspace learning-based methods [5, 26], nonnegative matrix factorization (NMF)-based methods [32, 9], graph learning-based methods [6, 11], and multiple kernel-based methods [16, 17]. These traditional shallow models often exhibit limited capabilities to conduct feature representation learning on large-scale datasets [31].

A number of deep learning-based methods have been proposed to alleviate the above problems [28, 37, 13, 21, 39, 42, 14, 35]. The goal of these deep MVC methods is to learn a more discriminative consensus representation by transforming each view with a corresponding view-specific encoder network. For example, Xie *et al.* [35] proposed a deep embedding-based clustering method that simultaneously learns feature representations and cluster assignments using deep neural networks. Li *et al.* [14] proposed a deep adversarial MVC method that learns the intrinsic structure embedded in multiview data. Zhou *et al.* [42] proposed an end-to-end adversarial attention network that makes use of adversarial learning and an attention mechanism to align latent feature distributions and evaluate the importance of different modalities. These methods yield significantly improved clustering performance. However, they fail to consider the semantic label consistency among multiple views, which may lead to difficulty in learning consistent cluster assignments.

Recently, contrastive learning has been integrated into deep learning models to learn discriminative representations of multiple views [12, 7]. Most existing contrastive learning-based methods attempt to maximize the mutual information contained among the assignment distributions of multiple views [37, 30, 3]. For example, Yang *et al.* [38] took advantage of the available data pairs as positive samples and randomly chose some cross-view samples as negative samples for MVC. In particular, the term “cross-view” means that any two views among multiple views are involved in the contrastive learning process. Caron *et al.* [3] presented an unsupervised visual feature learning method

*Corresponding author

that enforces consistency between the cluster assignments produced for different augmentations. Xu *et al.* [37] presented a multilevel feature learning (MFL) method to generate features at different levels for contrastive MVC, e.g., low-level features, high-level features and semantic features. However, which features play critical roles in contrastive feature learning remains unknown [27, 30]. This still leaves an important open question: "which representation should be invariant to multiple views?" Therefore, this motivates us to develop a cross-view contrastive learning (CVCL) model to build a more reasonable view-invariant representation scheme for multiview learning.

In this paper, we present a CVCL method that learns view-invariant representations for MVC. In contrast with most existing deep MVC methods, a cluster-level CVCL strategy is introduced to capture consistent semantic label information across multiple views. By contrasting cluster assignments among multiple views, the proposed CVCL method learns view-invariant representations between positive pairs for MVC. The cluster assignments derived from positive pairs reasonably align the invariant representations among multiple views. Such an alignment flexibly describes the consistent semantic labels obtained from individual views, which are used to measure the intrinsic relationships among the data samples. The K -dimensional assignment probability represents the cluster assignment of each sample in the corresponding view. Based on these view-invariant representations, the contrastive loss of the proposed CVCL method encourages the K -dimensional cluster assignments produced for positive pairs to be similar and pushes the cluster assignments provided for negative pairs apart. In addition, we provide a theoretical explanation for the realizability of soft cluster assignment alignment.

Our major contributions are summarized as follows.

- A CVCL model that contains a two-stage training scheme is proposed to learn view-invariant representations in an end-to-end manner.
- By contrasting the cluster assignments among multiple views, a cluster-level CVCL strategy is presented to explore consistent semantic label information.
- A theoretical analysis of the alignment among the produced view-invariant representations explains why the CVCL model is able to work effectively under certain conditions.
- Extensive experiments conducted on seven multiview datasets demonstrate the effectiveness of the proposed CVCL method.

2. Related Work

In this section, we briefly introduce some work related to the proposed CVCL method, including studies on MVC

and contrastive learning.

2.1. Deep Multiview Clustering

Inspired by recent advances in deep neural network techniques, deep clustering approaches consisting of multiple nonlinear transformations have been extensively studied [29, 37, 7, 40, 30, 3]. One of the representative deep MVC methods is based on deep autoencoders with different regularization terms. These deep autoencoder-based MVC methods aim to learn a consensus representation by minimizing the reconstruction error induced by instances of multiple views. For example, *et al.* [35] proposed a deep embedded clustering (DEC) method using deep neural networks. DEC first transforms the input sample into features with a nonlinear mapping:

$$f_{\theta} : \mathbf{X} \rightarrow \mathbf{Z} \quad (1)$$

where θ is a learnable parameter set. Then the Kullback–Leibler (KL) divergence between the soft assignment \mathbf{q}_i and the auxiliary distribution \mathbf{p}_i is defined as follows:

$$L = KL(\mathbf{P}||\mathbf{Q}) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (2)$$

The KL divergence loss is minimized to improve the cluster assignment and feature representation effects.

2.2. Contrastive Learning

Contrastive learning has recently achieved significant progress in self-supervised representation learning [7, 27, 30]. Contrastive learning-based methods are essentially dependent on a large number of distinct pairwise representation comparisons. Specifically, these methods attempt to maximize the similarities among positive pairs and simultaneously minimize those among negative pairs in a latent feature space. The positive pairs are constructed from the invariant representations of all multiview instances of the same sample. The negative pairs are obtained from the invariant representations of multiple views for different samples. For example, Chen *et al.* [7] presented a visual representation framework for contrastive learning, which maximizes the agreement between differently augmented views of the same example in the latent feature space. Wang *et al.* [30] investigated the two key properties of the loss function of contrastive learning, i.e., the alignment of features derived from positive pairs and the uniformity of the feature distribution induced on the hypersphere, which can be used to measure the resulting representation quality. These methods are capable of learning good representations based on data argumentation. However, it remains challenging to determine invariant representations for multiple views.

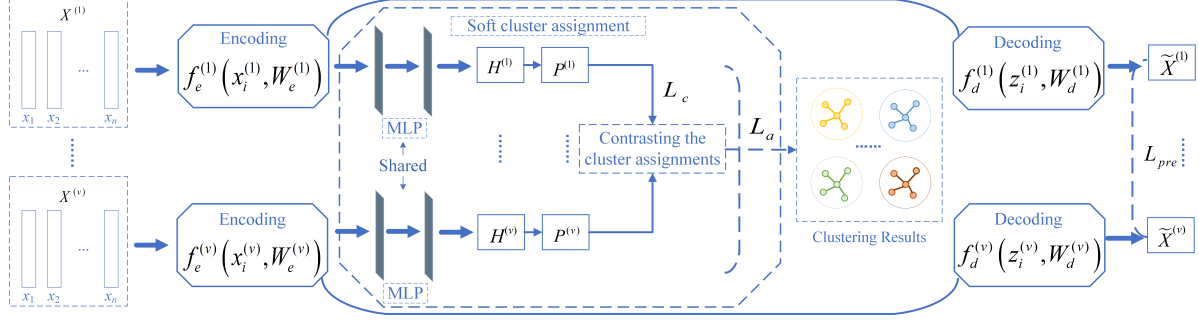


Figure 1. The framework of CVCL. Each view contains two modules, including a view-specific autoencoder module and a CVCL module. The multilayer perceptron (MLP) consists of multiple linear layers. The view-specific autoencoder module contains the encoding part and the decoding part, i.e., $\{f_e^{(v)}(\mathbf{x}_i^{(v)}, \mathbf{W}_e^{(v)})\}_{v=1}^{n_v}$ and $\{f_d^{(v)}(\mathbf{z}_i^{(v)}, \mathbf{W}_d^{(v)})\}_{v=1}^{n_v}$, respectively. The CVCL module is employed to explore consistent semantic label information by contrasting the cluster assignments among multiple views.

3. The Proposed Method

3.1. Proposed Statement

Given a set of multiview data $\mathcal{X} = \{\mathbf{X}^{(v)} \in \mathbb{R}^{d_v \times N}\}_{v=1}^{n_v}$ with n_v views and N samples, $\mathbf{X}^{(v)}$ represents the v th view of the multiview data. Each view $\mathbf{X}^{(v)} = [\mathbf{x}_1^{(v)}, \mathbf{x}_2^{(v)}, \dots, \mathbf{x}_N^{(v)}]$ has a total of N instances, where $\mathbf{x}_i^{(v)}$ ($1 \leq i \leq N$) represents a d_v -dimensional instance. Assume that K is the number of clusters. The samples with the same semantic labels can be grouped into the same cluster. Hence, N samples can be categorized into K different clusters.

3.2. Network Architecture

The goal of the proposed CVCL method is to produce semantic labels for end-to-end clustering from the raw instances of multiple views. We introduce an end-to-end deep clustering network architecture by applying contrastive cluster assignments to feature representation learning. As illustrated in Fig. 1, the proposed CVCL network architecture consists of two main modules, i.e., view-specific autoencoder module and cross-view contrastive learning module. The view-specific autoencoder module individually learns clustering-friendly features among multiple views under unsupervised representation learning. The cross-view contrastive learning module achieves the final cluster result by contrasting cluster assignments. With these two modules, CVCL simultaneously learns the view-invariant representations and produce the clustering result for MVC.

3.3. Cluster-Level CVCL

Let $f : \mathcal{X} \rightarrow \{\mathbf{Z}^{(v)} \in \mathbb{R}^{N \times k}\}_{v=1}^{n_v}$ be a function that maps N samples into semantic features. We stack two linear layers and a successive softmax function on the semantic features to produce a cluster assignment probability, which

is computed by

$$f\{\mathbf{W}_h^{(v)}\}_{v=1}^{n_v} : \{\mathbf{Z}^{(v)}\}_{v=1}^{n_v} \rightarrow \{\mathbf{H}^{(v)}\}_{v=1}^{n_v} \quad (3)$$

where $\{\mathbf{W}_h^{(v)}\}_{v=1}^{n_v}$ is a set of learnable parameters.

Inspired by recently proposed contrastive learning techniques, we employ these techniques on the semantic labels to explore the consistency information possessed across multiple views. We can obtain cluster probability matrices $\{\mathbf{H}^{(v)} \in \mathbb{R}^{N \times K}\}_{v=1}^{n_v}$ for all views, which are produced on the semantic features of the previous layer. Let $\mathbf{h}_i^{(v)}$ be the i th row in $\mathbf{H}^{(v)}$, and let $h_{ij}^{(v)}$ represent the probability that instance i belongs to cluster j in view m . The semantic label of instance i is determined by the largest value among the probabilities in $\mathbf{h}_i^{(v)}$.

To increase the differences among the cluster assignments, a unified target distribution $\{\mathbf{P}^{(v)} \in \mathbb{R}^{N \times K}\}_{v=1}^{n_v}$ is considered to be a good surrogate for $\{\mathbf{H}^{(v)}\}_{v=1}^{n_v}$, each element of which is calculated as follows:

$$p_{ij}^{(v)} = \frac{(\mathbf{h}_{ij}^{(v)})^2 / \sum_{i=1}^N \mathbf{h}_{ij}^{(v)}}{\sum_{k=1}^K \left((\mathbf{h}_{ik}^{(v)})^2 / \sum_{i=1}^N \mathbf{h}_{ik}^{(v)} \right)}. \quad (4)$$

Let $\mathbf{p}_j^{(v)}$ be the j th column of $\mathbf{P}^{(v)}$. Each element $p_{ij}^{(v)}$ in $\mathbf{p}_j^{(v)}$ indicates a soft cluster assignment of sample i belonging to cluster j . Thus, $\mathbf{p}_j^{(v)}$ represents a cluster assignment of the same semantic cluster.

The instances in the different views corresponding to an individual sample share common semantic information. The similarity between two cluster assignments $\mathbf{p}_j^{(v_1)}$ and $\mathbf{p}_j^{(v_2)}$ of cluster j is measured by

$$s(\mathbf{p}_j^{(v_1)}, \mathbf{p}_j^{(v_2)}) = (\mathbf{p}_j^{(v_1)})^T \mathbf{p}_j^{(v_2)} \quad (5)$$

where v_1 and v_2 denote two distinct views. The cluster assignment probabilities of the instances among different views should be similar in the CVCL module since these instances characterize the same sample. Moreover, the instances in multiple views are irrelevant to each other if they are used to characterize different samples. Therefore, there are $(n_v - 1)$ positive cluster assignment pairs and $n_v(K - 1)$ negative cluster assignment pairs when considering $\mathbf{p}_j^{(v)}$ and K clusters across n_v views.

The similarities among the intracluster assignments should be maximized, and those among the intercluster assignments should be minimized. We simultaneously cluster the samples while enforcing consistency among the cluster assignments. The cross-view contrastive loss between $\mathbf{p}_k^{(v_1)}$ and $\mathbf{p}_k^{(v_2)}$ is defined as follows:

$$l^{(v_1, v_2)} = -\frac{1}{K} \sum_{k=1}^K \log \frac{e^{s(\mathbf{p}_k^{(v_1)}, \mathbf{p}_k^{(v_2)})/\tau}}{T},$$

$$T = \sum_{j=1, j \neq k}^K e^{s(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_1)})/\tau} + \sum_{j=1}^K e^{s(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_2)})/\tau}$$
(6)

where τ is a temperature parameter, $(\mathbf{p}_k^{(v_1)}, \mathbf{p}_k^{(v_2)})$ is a positive cluster assignment pair between two views v_1 and v_2 , and $(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_1)})$ ($j \neq k$) and $(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_2)})$ are the negative cluster assignment pairs in two views v_1 and v_2 , respectively. The cross-view contrastive loss induced across multiple views is designed as:

$$L_c = \frac{1}{2} \sum_{v_1=1}^{n_v} \sum_{v_2=1, v_2 \neq v_1}^{n_v} l^{(v_1, v_2)}. \quad (7)$$

The cross-view contrastive loss explicitly compares pairs of cluster assignments among multiple views. It pulls pairs of cluster assignments from the same cluster together and pushes cluster assignments from different clusters away from each other.

To prevent all instances from being assigned to a particular cluster, we introduce a regularization term as follows:

$$L_a = \sum_{v=1}^{n_v} \sum_{j=1}^K q_j^{(v)} \log q_j^{(v)} \quad (8)$$

where $q_j^{(v)}$ is defined as $q_j^{(v)} = \frac{\sum_{i=1}^N p_{ij}^{(v)}}{N}$. This term is regarded as a cross-view consistency loss in the CVCL model [37]. Assume that all instances belong to a single cluster j . This implies that $p_{ij}^{(v)} = 1$ for all $i = 1, 2, \dots, N$ such that $q_j^{(v)} \log q_j^{(v)} = 0$. As $0 \leq p_{ij}^{(v)} \leq 1$, we have

$q_j^{(v)} \log q_j^{(v)} \leq 0$. This means that the following inequality,

$$q_j^{(v)} \ln q_j^{(v)} < 0, \quad (9)$$

holds if each cluster has at least one instance. This encourages more elements to reside in each greater-than-zero row of $\mathbf{P}^{(v)}$ by minimizing L_a in Eq. (8). Therefore, the network is able to encourage cross-view consistency across the cluster assignment probabilities of different instances among the multiple views of each sample with Eq. (8).

3.4. Two-Stage Training Scheme

As illustrated in Fig. 1, we first perform a pretraining task with a deep autoencoder for parameter initialization. Then, we employ a fine-tuning step to train the whole network for MVC.

3.4.1 Parameter Initialization via a Pretraining Network

We design a pretraining network that is made up of a view-specific encoder module $f_e^{(v)}$ ($1 \leq v \leq n_v$) and a corresponding decoder module $f_d^{(v)}$ for each view. The encoder module learns the embedded feature representations by

$$\mathbf{z}_i^{(v)} = f_e^{(v)}(\mathbf{x}_i^{(v)}, \mathbf{W}_e^{(v)}) \quad (10)$$

where $\mathbf{z}_i^{(v)}$ is the embedded feature representation of $\mathbf{x}_i^{(v)}$. The decoder module reconstructs the sample $\mathbf{x}_i^{(v)}$ as follows

$$\tilde{\mathbf{x}}_i^{(v)} = f_d^{(v)}(f_e^{(v)}(\mathbf{x}_i^{(v)}, \mathbf{W}_e^{(v)}), \mathbf{W}_d^{(v)}) \quad (11)$$

where $\tilde{\mathbf{x}}_i^{(v)}$ is a reconstruction of $\mathbf{x}_i^{(v)}$. Each encoder or decoder module consists of four or more layers in the proposed CVCL model. The nonlinear rectified linear unit (ReLU) function is chosen as the activation function in the deep autoencoder.

For multiple views, the reconstruction loss of the pretraining network between the input and output is designed as:

$$L_{pre} = \sum_{v=1}^{n_v} \sum_{i=1}^N \left\| \mathbf{x}_i^{(v)} - f_d^{(v)}(f_e^{(v)}(\mathbf{x}_i^{(v)}, \mathbf{W}_e^{(v)}), \mathbf{W}_d^{(v)}) \right\|^2 \quad (12)$$

This is considered to be a pretraining stage for parameter initialization.

3.4.2 MVC via a Fine-tuning Network

The overall loss of the proposed method consists of three main components: the reconstruction loss of the pretraining

Algorithm 1 Optimization procedure for CVCL

Input: Data matrices $\{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$, the numbers of samples N and epochs $epochs$, parameters α and β .

- 1: Initialize $\{\mathbf{W}^{(v)}\}_{v=1}^{n_v}$ by minimizing L_{pre} in Eq. (12);
- 2: **for** $t = 1$ to $epochs$ **do**
- 3: Choose a random minibatch of samples;
- 4: Computing $\{\tilde{\mathbf{X}}^{(v)}\}_{v=1}^{n_v}$ and $\{\mathbf{H}^{(v)}\}_{v=1}^{n_v}$ via Eqs. (10) and (3), respectively;
- 5: Computing $\{\mathbf{P}^{(v)}\}_{v=1}^{n_v}$ via Eq. (4);
- 6: Computing $\{\mathbf{W}^{(v)}\}_{v=1}^{n_v}$ and $\{\mathbf{W}_h^{(v)}\}_{v=1}^{n_v}$ by minimizing L_{fine} in Eq. (13);
- 7: **end for**
- 8: Calculate semantic labels by Eq. (14);

Output: The label predictions $\mathbf{Y} = [y_1, y_2, \dots, y_n]$.

network, the cross-view contrastive loss and the cross-view consistency loss, i.e.,

$$L_{fine} = L_{pre} + \alpha L_c + \beta L_a \quad (13)$$

where α and β are tradeoff hyperparameters.

The proposed method aims to learn the common semantic labels from the feature representations, which are generated from the instances of multiple views. Let $\mathbf{q}_i^{(v)}$ be the i th row of $\mathbf{P}^{(v)}$, and let $q_{ij}^{(v)}$ denote the j th element of $\mathbf{q}_i^{(v)}$. Specifically, $q_{ij}^{(v)}$ is the K -dimensional soft assignment probability, where $\sum_{i=1}^K q_{ij}^{(v)} = 1$. Once the training process of the network is completed, the semantic label of sample i ($1 \leq i \leq N$) can be predicted by

$$y_i = \arg \max_j \left(\frac{1}{n_v} \sum_{v=1}^{n_v} q_{ij}^{(v)} \right). \quad (14)$$

An adaptive momentum-based minibatch gradient descent method [22] is employed to optimize the whole network during the two training stages. The final clustering results are produced by the deep autoencoder with the CVCL module. The entire optimization procedure of the proposed method is summarized in Algorithm 1.

3.5. Theoretical Analysis

3.5.1 Generalization Bound of the Loss Function

We analyze the generalization bound of the loss function in the proposed method. According to Theorem 1, L_c has a specific lower bound in Eq. (7). The proof of Theorem 1 can be found in the supplementary material. Assume that each cluster must have at least one multiview data sample. This indicates that $q_{ij}^{(v)} > 0$ in $q_j^{(v)}$. A constant c must exist such that $q_{ij}^{(v)} \log q_{ij}^{(v)} > c$ in Eq. (8) holds for all

$j = 1, 2, \dots, K$. This shows that L_a also has a lower bound in Eq. (8). Therefore, a lower bound is theoretically guaranteed to be obtained when minimizing L_{fine} in Eq. (13).

Theorem 1 Assume that there are N samples and K clusters. Given two views v_1 and v_2 and $l^{(v_1, v_2)}$ in Eq. (6), the following inequality holds:

$$l^{(v_1, v_2)} \geq e^{\log(2K-1) - N/\tau}.$$

3.5.2 Realizability of Soft Cluster Assignment Alignment

For the sake of discussion, we assume that there are three clusters with sizes of k_1 , k_2 and k_3 and N samples ($N = k_1 + k_2 + k_3$). We consider an ideal case in which all instances in different clusters strictly belong to the respective low-dimensional subspaces in each view. Without loss of generality, each $\mathbf{P}^{(v)}$ can be represented by

$$\mathbf{P}^{(v)} = \begin{bmatrix} \mathbf{P}_1^{(v)} \\ \mathbf{P}_2^{(v)} \\ \mathbf{P}_3^{(v)} \end{bmatrix} = \begin{bmatrix} \overrightarrow{\mathbf{1}_{k_1}} & \overrightarrow{\mathbf{0}_{k_1}} & \overrightarrow{\mathbf{0}_{k_1}} \\ \overrightarrow{\mathbf{0}_{k_2}} & \overrightarrow{\mathbf{1}_{k_2}} & \overrightarrow{\mathbf{0}_{k_2}} \\ \overrightarrow{\mathbf{0}_{k_3}} & \overrightarrow{\mathbf{0}_{k_3}} & \overrightarrow{\mathbf{1}_{k_3}} \end{bmatrix} \quad (15)$$

where $\overrightarrow{\mathbf{1}_{k_1}}$ denotes a column vector of all ones with a size of k_1 . In particular, we always find a matrix transpose \mathbf{T} to obtain such a matrix $\mathbf{P}^{(v)}$ if the arrangement assumption is violated. For any two views v_1 and v_2 , $\mathbf{P}^{(v_1)}$ is identical to $\mathbf{P}^{(v_2)}$. Hence, the $\{\mathbf{P}^{(v)}\}_{v=1}^{n_v}$ are invariant to all types of instances for multiple views.

Definition 1 (Strict Alignment) Given an encoder f , we have $\mathbf{P}^{(v_1)} = f(\mathbf{X}^{(v_1)}) \in \mathbb{R}^{N \times K}$ and $\mathbf{P}^{(v_2)} = f(\mathbf{X}^{(v_2)}) \in \mathbb{R}^{N \times K}$, where $\mathbf{X}^{(v_1)}$ and $\mathbf{X}^{(v_2)}$ represent the instances of views v_1 and v_2 , respectively. The encoder f is strictly aligned if the following conditions are satisfied: $\forall v_1, v_2 \in \{1, 2, \dots, n_v\}, v_1 \neq v_2; \forall i, j \in \{1, 2, \dots, K\}, i \neq j$; and $\forall k \in \{1, 2, \dots, N\}$,

- (1) $\mathbf{p}_i^{(v_1)} = \mathbf{p}_i^{(v_2)}$;
- (2) $p_{ik}^{(v_1)} = \begin{cases} 1, & \text{the } k\text{th sample belongs to the } i\text{th cluster} \\ 0, & \text{otherwise} \end{cases}$;
- (3) $\langle \mathbf{p}_i^{(v_1)}, \mathbf{p}_j^{(v_1)} \rangle = \langle \mathbf{p}_i^{(v_1)}, \mathbf{p}_j^{(v_2)} \rangle = 0$

where $\mathbf{p}_i^{(v_1)}$ and $\mathbf{p}_i^{(v_2)}$ represent the i th columns of $\mathbf{P}^{(v_1)}$ and $\mathbf{P}^{(v_2)}$, respectively.

For any cluster assignment i ($1 \leq i \leq 3$) in $\mathbf{P}^{(v_1)}$, one positive cluster assignment pair and four negative cluster assignment pairs are produced for the two views v_1 and v_2 . To illustrate the realizability of similarity alignment, we introduce the definition of strict alignment.

Theorem 2 For n_v given views of multiview data, L_c in Eq. (7) is minimized if f is strictly aligned $\forall v_1, v_2 \in \{1, 2, \dots, n_v\}$ and $v_1 \neq v_2$.

Theorem 2 shows that a lower bound in Theorem 1 can be theoretically achieved when strict alignment is satisfied according to Definition 1. The proof of Theorem 2 is given in the supplementary material. Strict alignment is an ideal case, which implies that each cluster assignment i ($1 \leq i \leq 3$) in $\mathbf{P}^{(v)}$ ($1 \leq v \leq n_v$) satisfies the conditions in Definition 1. Specifically, the first condition affects a single positive cluster assignment pair while the other two conditions are imposed on the four negative cluster assignment pairs.

In the general case, designing an encoder that is strictly aligned for multiple views is an intractable problem. Let $\tilde{\mathbf{p}}_i^{(v)}$ be the i th row of $\mathbf{P}^{(v)}$, i.e., the feature of the i th instance in the v th view. Considering the cosine similarity measure, the distance between two features $\tilde{\mathbf{p}}_i^{(v_1)}$ and $\tilde{\mathbf{p}}_i^{(v_2)}$ is measured as

$$d\left(\tilde{\mathbf{p}}_i^{(v_1)}, \tilde{\mathbf{p}}_i^{(v_2)}\right) = \frac{\langle \tilde{\mathbf{p}}_i^{(v_1)}, \tilde{\mathbf{p}}_i^{(v_2)} \rangle}{\|\tilde{\mathbf{p}}_i^{(v_1)}\| \|\tilde{\mathbf{p}}_i^{(v_2)}\|} \quad (16)$$

where $\langle \cdot, \cdot \rangle$ is the dot product operator. The cosine similarity may be inaccurate when two instances of a sample in views v_1 and v_2 belong to different domains of multiview data, e.g., text and image pairs. In addition, the alignment sensitivity is insufficient when considering the similarity between two features $\tilde{\mathbf{p}}_i^{(v_1)}$ and $\tilde{\mathbf{p}}_i^{(v_2)}$ in Eq. (5). For the proposed CVCL method, $\mathbf{P}^{(v)}$ is theoretically invariant to all types of views. From the point of view of alignment, the alignment of the cluster assignments exhibits a stronger ability to perform invariant representation learning than that of the instance features in MVC.

3.5.3 Complexity Analysis

Let m and s denote the minibatch size and the maximum number of neurons in the hidden layers of the proposed network architecture, respectively. The complexity of the feedforward computation is $\mathcal{O}(n_v m d_v s^{(r+1)} + n_v m d_v K)$ in the fine-tuning phase. The complexities of the reconstruction loss, cross-view contrastive loss and cross-view consistency loss are $\mathcal{O}(n_v d_v m)$, $\mathcal{O}(m^2 K n_v ((n_v - 1) + n_v (K - 1)))$ and $\mathcal{O}(n_v K)$, respectively. Therefore, the overall complexity of the proposed CVCL method is $t(n_v m d_v s^{(r+1)} + n_v^2 m^2 K^2 + n_v m d_v K)$, where t is the maximum number of iterations in the pretraining and fine-tuning phases.

4. Experiments

In this section, we conduct extensive experiments to evaluate the performance of the proposed CVCL method. The source code for CVCL is implemented in Python 3.9. The source code is available at <https://github.com/chenjie20/CVCL>. All experiments are conducted on a Linux workstation with a GeForce RTX 2080 Ti GPU (11 GB caches), an Intel (R) Xeon (R) E5-2667 CPU and 256.0 GB of RAM.

4.1. Experimental Settings

4.1.1 Datasets

The proposed CVCL method is experimentally evaluated on seven publicly available multiview datasets. The MSRC-v1 dataset contains 210 scene recognition images belonging to 7 categories [33]. Each image is described by 5 different types of features. The COIL-20 dataset is composed of 1,440 images belonging to 20 categories [19]. Each image is described by 3 different types of features. The Handwritten dataset consists of 2,000 handwritten images of digits from 0 to 9 [1]. Each image is described by 6 different types of features. The BDGP dataset contains 2,500 samples of *Drosophila* embryos [2]. Each sample is represented by visual and textual features. The Scene-15 dataset consists of 4,485 scene images belonging to 15 classes [10]. Each image is represented by 3 different types of features. The MNIST-USPS dataset contains 5,000 samples with two different styles of digital images [1]. The Fashion dataset contains 10,000 images of products [34]. Each image is represented by three different styles.

4.1.2 Comparison Methods

To validate the superiority of the proposed CVCL method, we compare CVCL with several state-of-the-art methods, including the efficient and effective incomplete MVC (EE-IMVC) algorithm, augmented sparse representation (ASR) algorithm [6], deep safe IMVC (DSIMVC) algorithm [24], dual contrastive prediction (DCP) algorithm [15], deep safe MVC (DSMVC) algorithm [25] and MFL [37]. For DCP, the best clustering result is reported from the combinations of each pair of individual views in each dataset. In addition, two extra baselines are included for comparison purposes. Specifically, we first apply spectral clustering [18] on each individual view and report the best clustering result obtained among multiple views, i.e., the best single-view clustering (BSVC) method. Then, we apply an adaptive neighbor graph learning method [20] to produce a similarity matrix for each individual view. We aggregate all similarity matrices into an accumulated similarity matrix for spectral clustering, which is referred to as SC_{Agg} .

Table 1. Results of clustering performance comparisons conducted on all datasets.

Methods	MSRC-v1			COIL-20			Handwritten			BDGP			Scene-15			MNIST-USPS			Fashion		
	ACC	NMI	Purity	ACC	NMI	purity	ACC	NMI	purity	ACC	NMI	purity	ACC	NMI	purity	ACC	NMI	purity	ACC	NMI	purity
BSVC	78.57	68.04	78.57	80.21	84.75	80.47	75.35	74.07	75.35	53.68	32.42	54.32	38.05	38.85	42.08	67.98	74.43	72.34	60.32	64.91	63.84
SC _{Agg}	82.71	72.52	82.71	73.13	78.46	73.89	79.85	82.62	83.35	68	55.71	70.72	38.13	39.31	44.76	89	77.12	89.18	98	94.8	97.56
EE-IMVC	85.81	73.76	85.81	75.73	83.52	75.76	89.3	81.07	89.3	88	71.76	87.76	39	33.02	40.27	76	68.04	76.48	84	79.53	84.45
ASR	91.9	84.75	91.9	80.9	87.6	81.5	93.95	88.26	93.95	97.68	92.63	97.68	42.7	40.7	45.6	97.9	94.72	97.9	96.52	93.04	96.52
DSIMVC	79.05	69	79.05	65.55	72.51	66.67	87.2	80.39	87.2	99.04	96.86	99.04	28.27	29.04	29.79	99.34	98.13	99.34	88.21	83.99	88.21
DCP	78.57	74.84	79.43	67.36	78.79	69.86	85.75	85.05	85.75	97.04	92.43	97.04	42.32	40.38	43.85	99.02	97.29	99.02	89.37	88.61	89.37
DSMVC	85.24	76.96	85.24	76.46	84.15	78.19	96.8	92.57	96.8	75.8	61.39	75.8	43.48	41.11	45.92	96.34	94.27	96.34	89.63	86.81	89.63
MFL	93.33	86.51	93.33	73.19	81.43	75.07	86.55	85.98	86.55	98.72	96.13	98.72	42.52	40.34	44.53	99.66	99.01	99.66	99.2	98	99.2
CVCL	97.62	94.98	97.62	84.65	88.89	85.07	97.35	94.05	97.35	99.2	97.29	99.2	44.59	42.17	47.36	99.7	99.13	99.7	99.31	98.21	99.31

Table 2. Ablation study concerning the main components of the proposed CVCL method on all the datasets.

Methods	L_{pre}	L_c	L_a	MSRC-v1			COIL-20			Handwritten			BDGP			Scene-15			MNIST-USPS			Fashion		
				ACC	NMI	purity	ACC	NMI	purity	ACC	NMI	purity	ACC	NMI	purity	ACC	NMI	purity	ACC	NMI	purity	ACC	NMI	purity
CVCL _{fine-tuning}		✓	✓	82.38	76.17	82.38	66.94	78.96	69.86	88	87.87	88	99.12	96.88	99.12	40.42	39.8	43.81	99.48	98.5	99.48	99.16	97.87	99.16
CVCL _{Lc}	✓	✓		55.34	50.71	55.34	51.11	70.61	51.39	87.35	90.29	87.7	98.92	96.3	98.92	25.04	30.91	25.04	59.84	81.57	59.84	99.2	97.98	99.2
CVCL	✓	✓	✓	97.62	94.98	97.62	84.65	88.89	85.07	97.35	94.05	97.35	99.2	97.29	99.2	44.59	42.17	47.36	99.7	99.13	99.7	99.31	98.21	99.31

4.1.3 Evaluation Metrics

Three widely used metrics are employed to evaluate the clustering performance of all competing algorithms, including the clustering accuracy (ACC), normalized mutual information (NMI), and purity [4]. For example, ACC considers the best matching result between two assignments, i.e., a cluster assignment obtained from an MVC algorithm and a known ground-truth assignment. For these metrics, a larger value indicates better clustering performance.

4.1.4 Network Architecture and Parameter Settings

The proposed network architecture consists of an input layer, hidden layers in the view-specific encoders, an extra linear layer and a softmax layer. A set of view-specific autoencoders is contained in the pretraining stage. The number of hidden layers possessed by the view-specific encoders or decoders ranges from 3 to 5. For example, the sizes of the 5 hidden layers are set to $[d_v, 256, 512, 1024, 2048, r_1]$, where d_v is the dimensionality of an instance in the v th view and r_1 is the dimensionality of the corresponding feature. Moreover, r_2 represents the size of the extra linear layer. We choose r_1 and r_2 from $[2000, 1500, 1024, 1000, 768, 512, 500, 256, 200]$. The overall loss of the proposed network architecture has two parameters, α and β , which are chosen from $\{0.005, 0.01, 0.05\}$ with a grid search strategy. For a fair comparison, the best clustering results of these competing methods are obtained by tuning their parameters.

4.2. Performance Evaluation

The clustering results produced by all competing methods on the seven multiview datasets are reported in Table 1. The best and second-best values of the clustering results are highlighted in bold and underlined, respectively. The contrastive learning-based methods, including CVCL, DSMVC, DSIMVC and MFL, often achieve significant improvements over the other methods on large-scale datasets, e.g., the BDGP, MNIST-USPS, Scene-15

and Fashion datasets. Moreover, the CVCL method significantly outperforms the other contrastive learning-based methods, including DSMVC, DCP, DSIMVC and MFL, on all datasets. This verifies the importance of the cluster-level CVCL strategy. These results demonstrate the effectiveness of our proposed CVCL method. The proposed CVCL method achieves the best clustering results on all datasets. This is consistent with our theoretical analysis. For example, the CVCL method achieves performance improvements of approximately 4.29%, 8.47%, and 4.29% over the second-best method on the MSRC-v1 dataset in terms of the ACC, NMI, and purity metrics, respectively. Similarly, the CVCL method performs much better than the other competing methods on the other datasets. These results demonstrate the superiority of CVCL over the other methods.

Two reasons explain the advantages and effectiveness of the proposed CVCL method. First, contrastive learning-based methods, e.g., CVCL, DSMVC, DSIMVC and MFL, consider deep representations for multiple views. We observe that they often achieve significant improvements over the other traditional methods, especially on large-scale datasets. Second, the alignment of soft cluster assignments plays a critical role in contrastive learning. By contrasting the cluster assignments among multiple views, the proposed CVCL method is guided to learn view-invariant representations in unsupervised learning. Consequently, it learns more discriminative view-invariant representations than the other contrastive learning-based methods.

4.3. Ablation Studies

According to the overall reconstruction loss in Eq. (13), three different loss components are included. To verify the importance of each component in CVCL, we perform ablation studies with the same experimental settings to isolate the necessity of each component. Specifically, we consider two special cases: performing MVC in the fine-tuning stage without pretraining and performing MVC in both stages without the regularization term of the overall reconstruction loss L_a . These versions are referred to as CVCL_{fine-tuning}

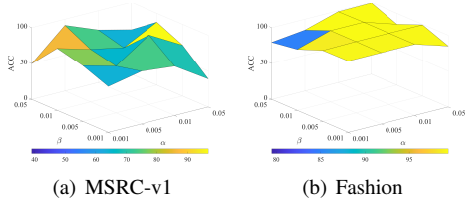


Figure 2. The ACC values yielded by the CVCL method with different α and β combinations on the four representative datasets.

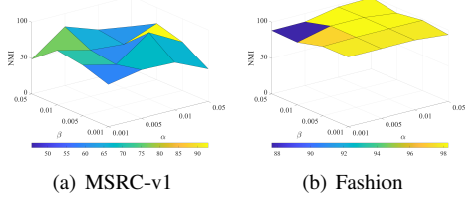


Figure 3. The NMI values yielded by the CVCL method with different α and β combinations on the two representative datasets.

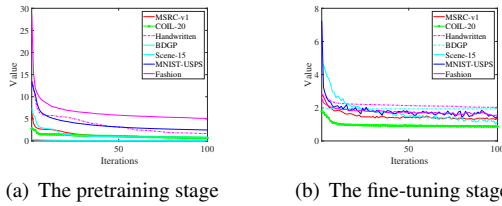


Figure 4. Convergence results obtained by the CVCL method on all the datasets.

and $CVCL_{L_c}$, respectively.

Table 2 shows the obtained clustering results in terms of the three metrics produced with the combinations of different loss components. The clustering results in the first two rows of Table 2 are achieved by the two special cases. As expected, the best performance can be achieved when all loss terms are considered and when the two-stage training scheme is employed in CVCL. Moreover, we can observe that the clustering performance is significantly improved when the pretraining stage is employed in CVCL. For example, CVCL performs much better than $CVCL_{fine-tuning}$, with improvements of approximately 5.24%, 18.81% and 15.24% in terms of the ACC, NMI and purity metrics, respectively, achieved on the MSRC-v1 dataset. However, the clustering performance gap narrows as the number of samples significantly increases. For example, CVCL achieves 0.08%, 0.22% and 0.15% ACC improvements over $CVCL_{fine-tuning}$ on the BDGP, MNIST-USPS and Fashion datasets, respectively. This indicates that an increase in the number of samples may reduce the significant advantages provided by the pretraining stage. In addition, the clustering performance achieved on most datasets dramatically declines when L_a is ignored in the overall reconstruction loss. This indicates that effectively guarantees that all instances can be assigned into clusters. Therefore, each component in the overall reconstruction loss plays a crucial role in learning view-invariant representations.

4.4. Parameter Sensitivity Analysis

We conduct experiments on two representative datasets, i.e., the MSRC-v1 and Fashion datasets, to investigate the sensitivity of the α and β parameters in the proposed CVCL method. The α and β parameters are chosen from $\{0.001, 0.005, 0.01, 0.05\}$ for CVCL. Figures 2 and 3 show the clustering performance achieved by the CVCL method in terms of the ACC and NMI values obtained with different combinations of α and β . It can be observed that the clustering performance attained by the CVCL method on the MSRC-v1 dataset seriously fluctuates with different combinations of α and β . As the number of samples dramatically increases in the other dataset, the CVCL method can achieve relatively stable clustering results with most combinations of α and β . This indicates that the CVCL method has stable clustering performance when utilizing a larger number of samples.

4.5. Training Analysis

We investigate the convergence of the CVCL method. Two major learning stages are contained in the CVCL method, including the pretraining and fine-tuning stages. To validate the convergence of the CVCL method, we compute the results of the loss functions in Eqs. (12) and (13) during these two stages. Figure 4 shows the curves of the loss function results obtained on all the datasets. The values of the loss function in Eq. (12) dramatically drop in the first few iterations and then slowly decrease until convergence is achieved. We also observe a similar trend in the changes in the loss function values in Eq. (13) on most datasets, e.g., the COIL-20, Handwritten and BDGP datasets. In addition, the curves of the loss function in Eq. (13) produced on the other datasets fluctuate slightly after the first few iterations. These results demonstrate the effectiveness of the convergence property of the CVCL method.

5. Conclusion

In this paper, we propose a CVCL method that learns view-invariant representations for MVC. A cluster-level CVCL strategy is presented to explore the consistent semantic label information possessed among multiple views. CVCL effectively achieves more discriminative cluster assignments during two successive stages. A theoretical analysis of soft cluster assignment alignment indicates the importance of the cluster-level learning strategy in CVCL. We conduct extensive experiments and ablation studies on MVC datasets to validate the superiority of the model and the effectiveness of each component in the overall reconstruction loss.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 62176171 and Grant U21B2040.

References

- [1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.
- [2] X. Cai, H. Wang, H. Huang, and C. Ding. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12):i16–i24, Jun. 2012.
- [3] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *in Proc. 34th Adv. Neural Inf. Process. Syst.*, pages 9912–9924, Vancouver, Canada, Dec. 2020.
- [4] J. Chen, H. Mao, Z. Wang, and X. Zhang. Low-rank representation with adaptive dictionary learning for subspace clustering. *Knowl. Based Syst.*, 22(8):107053, Jul. 2021.
- [5] J. Chen, S. Yang, H. Mao, and C. Fahy. Multiview subspace clustering using low-rank representation. *IEEE Trans. Cybern.*, 52(11):12364–12378, Nov. 2022.
- [6] J. Chen, S. Yang, X. Peng, D. Peng, and Z. Wang. Augmented sparse representation for incomplete multiview clustering. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1–14, Sept. 2022.
- [7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *in Proc. 37th Int. Conf. Mach. Learn.*, pages 1597–1607, Jul. 2020.
- [8] Z. Han, C. Zhang, H. Fu, and J. T. Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Trans. Pattern Anal. and Mach. Intell.*, pages 1–16, May 2022.
- [9] M. Hu and S. Chen. One-pass incomplete multi-view clustering. In *in Proc. 32th AAAI Conf. Artif. Intell.*, pages 3838–3845, Jan. 2019.
- [10] L. Fei-Fei L and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 524–531, San Diego, CA, USA, Jun. 2005.
- [11] L. Li, Z. Wan, and H. He. Incomplete multi-view clustering with joint partition and graph learning. *IEEE Trans. Knowl. Data Eng.*, pages 1–15, May 2021.
- [12] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng. Contrastive clustering. In *in Proc. 34th AAAI Conf. Artif. Intell.*, pages 8547–8555, Jan. 2021.
- [13] Y. Li, M. Yang, D. Peng, T. Li, J. Huang, and X. Peng. Twin contrastive learning for online clustering. *Int. J. Comput. Vis.*, 130:2205–2221, Jul. 2022.
- [14] Z. Li, Q. Wang, Z. Tao, Q. Gao, and Z. Yang. Deep adversarial multi-view clustering network. In *in Proc. 28th Int. Joint Conf. Artif. Intell.*, pages 2952–2958, Jul. 2019.
- [15] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Trans. Pattern Anal. and Mach. Intell.*, pages 1–14, Aug. 2022.
- [16] X. Liu, M. Li, C. Tang, J. Xia, J. Xiong, L. Liu, M. Kloft, and E. Zhu. Efficient and effective regularized incomplete multi-view clustering. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 43(8):2634–2646, Aug. 2021.
- [17] X. Liu, X. Zhu, M. Li, L. Wang, C. Tang, J. Yin, D. Shen, H. Wang, and W. Gao. Late fusion incomplete multi-view clustering. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 41(10):2410–2423, Oct. 2018.
- [18] U. Von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, Aug. 2007.
- [19] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-20). Technical Report Technical Report CUCS-005-96, Feb. 1996.
- [20] F. Nie, X. Wang, and H. Huang. Clustering and projected clustering with adaptive neighbors. In *in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pages 977–986, New York, USA, Aug. 2014.
- [21] X. Peng, Y. Li, I. W. Tsang, J. Lv H. Zhu, and J. T. Zhou. XAI beyond classification: interpretable neural clustering. *J. Mach. Learn. Res.*, 23(6):1–28, Jul. 2022.
- [22] N. Qian. On the momentum term in gradient descent learning algorithms. *Neural Netw.*, 12(1):145–151, Jan. 1999.
- [23] C. Tang, Z. Li, J. Wang, X. Liu, W. Zhang, and E. Zhu. Unified one-step multi-view spectral clustering. *IEEE Trans. Knowl. Data Eng.*, pages 1–11, May 2022.
- [24] H. Tang and Y. Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *in Proc. 39th Int. Conf. Mach. Learn.*, pages 21090–21110, Jul. 2022.
- [25] H. Tang and Y. Liu. Deep safe multi-view clustering: reducing the risk of clustering performance degradation caused by view increase. In *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 202–211, New Orleans, Louisiana, USA, Jun. 2022.
- [26] Z. Tao, J. Li, H. Fu, Y. Kong, and Y. Fu. From ensemble clustering to subspace clustering: Cluster structure encoding. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1–12, Sept. 2021.
- [27] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning. In *in Proc. 34th Adv. Neural Inf. Process. Syst.*, pages 6827–6839, Vancouver, Canada, Dec. 2020.
- [28] Q. Wang, Z. Tao, W. Xia, Q. Gao, X. Cao, and L. Jiao. Adversarial multiview clustering networks with adaptive fusion. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1–13, Feb. 2022.
- [29] Q. Wang, Z. Tao, W. Xia, Q. Gao, X. Cao, and L. Jiao. Adversarial multiview clustering networks with adaptive fusion. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1–13, Feb. 2022.
- [30] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *in Proc. 37th Int. Conf. Mach. Learn.*, pages 9929–9939, Jul. 2020.
- [31] Y. Wang, D. Chang, Z. Fu, J. Wen, and Y. Zhao. Graph contrastive partial multi-view clustering. *IEEE Trans. Multimedia*, pages 1–12, Sept. 2022.

- [32] Z. Wei, C. Xu, Z. Guan, and Y. Liu. Multiview concept learning via deep matrix factorization. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(2):814–825, Feb. 2021.
- [33] J. Winn and N. Jojic. Locus: learning object classes with unsupervised segmentation. In *in 10th IEEE Int. Conf. Comput. Vis.*, pages 756–763, Beijing, China, Oct. 2005.
- [34] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- [35] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *in Proc. 33th Int. Conf. Mach. Learn.*, pages 478–487, Jul. 2016.
- [36] J. Xu, T. Ren, H. Tang, Z. Yang, L. Pan, Y. Yang, X. Pu, P. S. Yu, and L. He. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Trans. Knowl. Data Eng.*, pages 1–12, Jul. 2022.
- [37] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, and L. He. Multi-level feature learning for contrastive multi-view clustering. In *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 16051–16060, New Orleans, Louisiana, USA, Jun. 2022.
- [38] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 14308–14317, New Orleans, USA, Jun. 2022.
- [39] M. Yang, Y. Li, P. Hu, J. Bai, J. Lv, and X. Peng. Robust multi-view clustering with incomplete information. *IEEE Trans. Pattern Anal. and Mach. Intell.*, pages 1–14, Mar. 2022.
- [40] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu. Deep partial multi-view learning. *IEEE Trans. Pattern Anal. and Mach. Intell.*, pages 1–14, Nov. 2020.
- [41] P. Zhang, X. Li, S. Zhou, W. Zhao, and E. Zhu Z. Cai. Consensus one-step multi-view subspace clustering. *IEEE Trans. Knowl. Data Eng.*, 31(10):4676–4689, Oct. 2022.
- [42] R. Zhou and Y.-D. Shen. End-to-end adversarial-attention network for multi-modal clustering. In *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, page 14619–14628, New Orleans, USA, Jun. 2020.

Deep Multiview Clustering by Contrasting Cluster Assignments —Supplementary Material

Jie Chen¹, Hua Mao², Wai Lok Woo², Xi Peng^{1*}

¹ College of Computer Science, Sichuan University, China

² Department of Computer and Information Sciences, Northumbria University

chenjie2010@scu.edu.cn; {hua.mao, wailok.woo}@northumbria.ac.uk;

pengx.gm@gmail.com

1. Additional Experiments

Table 1. Computation times (in seconds) of the other contrastive learning-based methods on all the datasets.

Methods	MSRC-v1	COIL-20	Handwritten	BDGP	Scene-15	MNIST-USPS	Fashion
DSIMVC	676.78	480.59	2368.74	1234.33	1770.68	4236.72	2368.74
DCP	106.69	158.73	265.39	219.8	626.42	509.07	785.25
DSMVC	261.74	886.39	802.22	865.46	1233.89	1164.54	4130.62
MFL	123.81	496.78	688.11	68.31	1430.29	511.67	939.47
CVCL	97.55	149.05	235.69	41.91	605.12	449.78	687.03

1.1. Investigating the Computational Costs

We compare the proposed CVCL method with the other contrastive learning-based methods in terms of their computational costs. With the enhanced learning capabilities, the importance of the computational cost may become secondary to the improved performance achievable by contrastive learning-based methods. Table 1 shows the running times of all the competing algorithms on all the datasets. It is clear that CVCL performs more efficiently than the other algorithms. This demonstrates the advantages of the proposed CVCL method in terms of computational efficiency.

1.2. Discussion

The instances of a sample from different views may sit on different underlying distributions. This means that the contrastive learning of the high-level and low-level features may not be reasonable in MFL [1]. For a given sample, the results of cluster assignments of its instances from multiple views trend to be consistent in CVCL. In contrast with MFL, CVCL ensures consistency among the cluster assignments produced from multiple views. The semantic label of each sample can be predicted using Eq. (14). Moreover, we provide a theoretical analysis for soft cluster assignment alignment. This explains why CVCL performs significantly better than MFL on some of the datasets.

*Corresponding author

2. Detailed Proofs

2.1. Proof of Theorem 1

Theorem 1 Assume that there are N samples and K clusters. Given two views v_1 and v_2 and $l^{(v_1, v_2)}$ in Eq. (6), the following inequality holds:

$$l^{(v_1, v_2)} \geq e^{\log(2K-1) - N/\tau}.$$

Proof Let $\mathbf{p}_j^{(v_1)}$ and $\mathbf{p}_j^{(v_2)}$ be the j th columns of $\mathbf{P}^{(v_1)}$ and $\mathbf{P}^{(v_2)}$, respectively. The i th elements $p_{ij}^{(v_1)}$ and $p_{ij}^{(v_2)}$ in $\mathbf{p}_j^{(v_1)}$ and $\mathbf{p}_j^{(v_2)}$ represent the cluster assignment probabilities, i.e., $0 \leq p_{ij}^{(v_1)} \leq 1$ and $0 \leq p_{ij}^{(v_2)} \leq 1$, respectively, where $1 \leq i \leq N$. Thus, we have

$$0 \leq s(\mathbf{p}_j^{(v_1)}, \mathbf{p}_j^{(v_2)}) \leq N \quad \text{and} \quad e^{s(\mathbf{p}_j^{(v_1)}, \mathbf{p}_j^{(v_2)})} \geq 1.$$

Suppose that

$$l = \frac{e^{s(\mathbf{p}_k^{(v_1)}, \mathbf{p}_k^{(v_2)})/\tau}}{\sum_{j=1, j \neq k}^K e^{s(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_1)})/\tau} + \sum_{j=1}^K e^{s(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_2)})/\tau}}$$

and we obtain

$$\begin{aligned} \log l &= \log \left(\frac{e^{s(\mathbf{p}_k^{(v_1)}, \mathbf{p}_k^{(v_2)})/\tau}}{\sum_{j=1, j \neq k}^K e^{s(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_1)})/\tau} + \sum_{j=1}^K e^{s(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_2)})/\tau}} \right) \\ &\quad - s(\mathbf{p}_k^{(v_1)}, \mathbf{p}_k^{(v_2)})/\tau \\ &\geq \log(2K-1) - N/\tau. \end{aligned}$$

Hence,

$$l^{(v_1, v_2)} \geq e^{\log(2K-1) - N/\tau}.$$

□

2.2. Proof of Theorem 2

Theorem 2 For n_v given views of multiview data, L_c in Eq. (7) is minimized if f is strictly aligned $\forall v_1, v_2 \in \{1, 2, \dots, n_v\}$ and $v_1 \neq v_2$.

Proof According to $\mathbf{p}_i^{(v_1)} = \mathbf{p}_i^{(v_2)}$ and the result of $p_{ik}^{(v_1)}$, we obtain

$$s\left(\mathbf{p}_i^{(v_1)}, \mathbf{p}_i^{(v_2)}\right) = \left(\mathbf{p}_i^{(v_1)}\right)^T \mathbf{p}_i^{(v_2)} = k_i$$

where k_i equals the number of samples in the i th cluster. Similarly,

$$s\left(\mathbf{p}_i^{(v_1)}, \mathbf{p}_j^{(v_1)}\right) = s\left(\mathbf{p}_i^{(v_1)}, \mathbf{p}_j^{(v_2)}\right) = 0.$$

Hence,

$$l^{(v_1, v_2)} = e^{\log(2K-1) - N/\tau}.$$

This shows that L_c in Eq. (7) is minimized. \square

References

- [1] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, and L. He. Multi-level feature learning for contrastive multi-view clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 16051–16060, New Orleans, Louisiana, USA, Jun. 2022. 1