Multi-level Feature Learning for Contrastive Multi-view Clustering

Jie Xu[1†], Huayi Tang[1†], Yazhou Ren[1*], Liang Peng[1], Xiaofeng Zhu[1,2], Lifang He[3]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

[2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518000, China

[3]Department of Computer Science and Engineering, Lehigh University, PA 18015, USA

jiexuwork@outlook.com,tangh4681@gmail.com,yazhou.ren@uestc.edu.cn

larrypengliang@gmail.com,seanzhuxf@gmail.com,lih319@lehigh.edu

Abstract

Multi-view clustering can explore common semantics from multiple views and has attracted increasing attention. However, existing works punish multiple objectives in the same feature space, where they ignore the conflict between learning consistent common semantics and reconstructing inconsistent view-private information. In this paper, we propose a new framework of multi-level feature learning for contrastive multi-view clustering to address the aforementioned issue. Our method learns different levels of features from the raw features, including low-level features, high-level features, and semantic labels/features in a fusion-free manner, so that it can effectively achieve the reconstruction objective and the consistency objectives in different feature spaces. Specifically, the reconstruction objective is conducted on the low-level features. Two consistency objectives based on contrastive learning are conducted on the high-level features and the semantic labels, respectively. They make the high-level features effectively explore the common semantics and the semantic labels achieve the multi-view clustering. As a result, the proposed framework can reduce the adverse influence of view-private information. Extensive experiments on public datasets demonstrate that our method achieves state-of-the-art clustering effectiveness.

1. Introduction

Multi-view clustering (MVC) is attracting more and more attention in recent years [22, 50, 52, 57] as multi-view data or multi-modal data can provide common semantics to improve the learning effectiveness [3, 14, 27, 33, 36, 43]. In the literature, existing MVC methods can be roughly divided into two categories, *i.e.*, traditional methods and deep methods.

The traditional MVC methods conduct the clustering task based on traditional machine learning methods and can be subdivided into three subgroups, including subspace methods [6, 18, 24], matrix factorization methods [45, 53, 56], and graph methods [28, 55, 60]. Many traditional MVC methods have the drawbacks such as poor representation ability and high computation complexity, resulting in limited performance in the complex scenarios with real-world data [10].

Recently, deep MVC methods have gradually become a popular trend in the community due to the outstanding representation ability [1, 2, 20, 44, 49, 50, 54]. Previous deep MVC methods can be subdivided into two subgroups, *i.e.*, two-stage methods and one-stage methods. Two-stage deep MVC methods (*e.g.*, [21, 50]) focus on separately learning the salient features from multiple views and performing the clustering task. However, Xie *et al*. [48] present that the clustering results can be leveraged to improve the quality of feature learning. Therefore, one-stage deep MVC methods (*e.g.*, [39, 59]) embed the feature learning with the clustering task in a unified framework to achieve end-to-end clustering.

Multi-view data contains two kinds of information, *i.e.*, the common semantics across all views and the view-private information for individual view. For example, a text and an image can be combined to describe common semantics, while the unrelated context in the text and the background pixels in the image are meaningless view-private information for learning common semantics. In multi-view learning, it is an always-on topic to learn common semantics and avoid the misleading of meaningless view-private information. Although important progress has been achieved by existing MVC methods, they have the following drawbacks to be addressed: (1) Many MVC methods (*e.g.*, [39, 59]) try to discover the latent cluster patterns by fusing the features of all views. However, the meaningless view-private information might be dominant in the feature fusion process, compared to the common semantics, and thus interferes with the quality of clustering. (2) Some MVC methods (*e.g.*, [18, 21]) leverage the consistency objective on the latent features to explore the common semantics across all views. However, they usually need the reconstruction objective on the same

[†]Equal contribution. [*]Corresponding author.

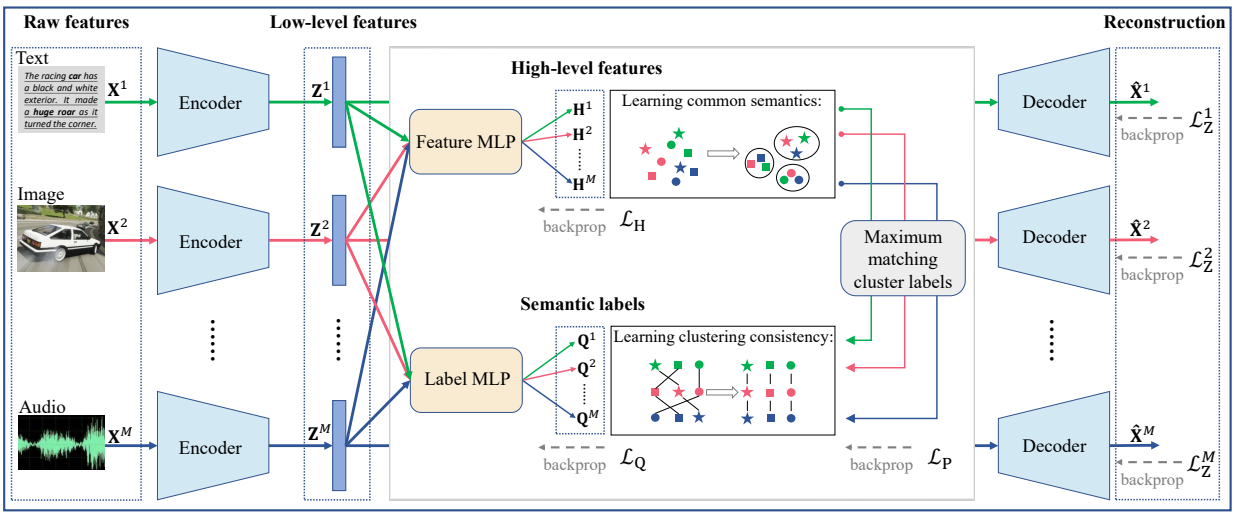


Figure 1. The framework of MFLVC. We avoid the direct feature fusion in multi-level feature learning, which learns the low-level features $\mathbf{Z}^m$, the high-level features $\mathbf{H}^m$, and the semantic labels $\mathbf{Q}^m$ from the raw features $\mathbf{X}^m$ for each view. The reconstruction objective $\mathcal{L}_{\mathbf{Z}}^m$ is individually conducted on $\mathbf{Z}^m$. Two consistency objectives (*i.e.*, $\mathcal{L}_{\mathbf{H}}$ and $\mathcal{L}_{\mathbf{Q}}$) are conducted on $\{\mathbf{H}^m\}_{m=1}^M$ and $\{\mathbf{Q}^m\}_{m=1}^M$, respectively. Furthermore, $\mathcal{L}_{\mathbf{P}}$ is optimized to leverage the cluster information of $\{\mathbf{H}^m\}_{m=1}^M$ to improve the clustering effectiveness of $\{\mathbf{Q}^m\}_{m=1}^M$.

features to avoid the trivial solution. This induces the conflict that the consistency objective tries to learn the features with common semantics across all views as much as possible while the reconstruction objective hopes the same features to maintain the view-private information for individual view.

In this paper, we propose a new framework of multi-level feature learning for contrastive multi-view clustering (MFLVC for short) to address the aforementioned issues, as shown in Figure 1. Our goals include (1) designing a fusion-free MVC model to avoid fusing the adverse view-private information among all views and (2) generating different levels of features for the samples in each view including low-level features, high-level features, and semantic labels/features. To do this, we first leverage the autoencoder to learn the low-level features from raw features, and then obtain the high-level features and semantic labels via stacking two MLPs on low-level features. Each MLP is shared by all views and is conducive to filtering out the view-private information. Furthermore, we take the semantic labels as the anchors, which combine with the cluster information in the high-level features to improve the clustering effectiveness. In this framework, the reconstruction objective is achieved by the low-level features while two consistency objectives are achieved by the high-level features and the semantic labels, respectively. Moreover, these two consistency objectives are conducted by contrastive learning, which makes the high-level features focus on mining the common semantics across all views and makes the semantic labels represent consistent cluster labels for multi-view clustering, respectively. As a result, the conflict between the reconstruction objective and two consistency objectives is alleviated. Compared to previous works, our contributions are listed as follows:

- We design a fusion-free MVC method which conducts different objectives in different feature spaces to solve the conflict between the reconstruction and consistency objectives. In this way, our method is able to effectively explore the common semantics across all views and avoid their meaningless view-private information.
- We propose a flexible multi-view contrastive learning framework, which can be used to simultaneously achieve the consistency objectives for the high-level features and the semantic labels. The high-level features enjoy good manifolds and represent common semantics, which enable to improve the quality of semantic labels.
- Our method is robust to the hyper-parameters' setting due to the well-designed framework. We conduct ablation studies in details, including the loss components and contrastive learning structures to understand the proposed model. Extensive experiments demonstrate that it achieves state-of-the-art clustering effectiveness.

2. Related Work

Multi-view clustering. The first category of MVC methods belongs to subspace clustering [18, 24], which focuses on learning a common subspace representation for multiple views. For instance, the traditional subspace clustering was extended by [6], where the authors presented a diversity-induced mechanism for multi-view subspace clustering. The second category of MVC methods is based on the matrix factorization technique [23, 56] that is formally equivalent to

the relaxation of K-means [26]. For example, Cai *et al*. [4] introduced a shared clustering indicator matrix for multiple views and handled a constrained matrix factorization problem. The third category of MVC methods is graph based MVC [28, 34], where graph structures are build to preserve the adjacency relationship among samples. The fourth category of MVC methods is based on deep learning framework, as known as deep MVC methods, which have been exploited increasingly and can be further roughly divided into two groups, *i.e*., two-stage deep MVC methods [21, 50] and one-stage deep MVC methods [20, 51, 59]. These methods utilize the excellent representation ability of deep neural networks to discover the latent cluster patterns of multi-view data.

Contrastive learning. Contrastive learning [7, 42] is an attention-getting unsupervised representation learning method, with the idea that maximizing the similarities of positive pairs while minimizing that of negative pairs in a feature space. This learning paradigm has lately achieved promising performance in computer vision, such as [29, 40]. For example, a one-stage online image clustering method was proposed in [19], which explicitly conducted contrastive learning in the instance-level and cluster-level. For multi-view learning, there are also some works based on contrastive learning [12, 21, 35, 38]. For instance, Tian *et al*. [38] proposed a contrastive multi-view coding framework to capture underlying scene semantics. In [12], the authors developed a multi-view representation learning method to tackle graph classification via contrastive learning. Recently, some works investigated different contrastive learning frameworks for multi-view clustering [21, 31, 39].

3. Method

Raw features. A multi-view dataset $\{\mathbf{X}^m \in \mathbb{R}^{N\times D_m}\}_{m=1}^M$ includes N samples across M views, where $\mathbf{x}_i^m \in \mathbb{R}^{D_m}$ denotes the D_m-dimensional sample from the m-th view. The dataset is treated as the raw features where multiple views have K common cluster patterns to be discovered.

3.1. Motivation

The multi-view data usually have redundancy and random noise, so the mainstream methods always learn salient representations from raw features. In particular, autoencoder [13, 37] is a widely used unsupervised model and it can project the raw features into a customizable feature space. Specifically, for the m-th view, we denote $E^m(\mathbf{X}^m;\theta^m)$ and $D^m(\mathbf{Z}^m;\phi^m)$, respectively, as the encoder and the decoder, where θ^m and ϕ^m are network parameters, denote $\mathbf{z}_i^m = E^m(\mathbf{x}_i^m) \in \mathbb{R}^L$ as the L-dimensional latent feature of the i-th sample, and denote $\mathcal{L}_{\mathbf{Z}}^m$ as the reconstruction loss between input $\mathbf{X}^m$ and output $\hat{\mathbf{X}}^m \in \mathbb{R}^{N\times D_m}$, so the reconstruction objective of all views is formulated as:

$$\mathcal{L}_{\mathbf{Z}} = \sum_{m=1}^{M} \mathcal{L}_{\mathbf{Z}}^m = \sum_{m=1}^{M}\sum_{i=1}^{N} \|\mathbf{x}_i^m - D^m(E^m(\mathbf{x}_i^m))\|_2^2 . \quad (1)$$

Based on $\{\mathbf{Z}^m = E^m(\mathbf{X}^m)\}_{m=1}^M$, MVC aims to mine the common semantics across all views to improve the clustering quality. To achieve this, existing MVC methods still have two challenges to be addressed: (1) Many MVC methods (*e.g*., [20, 59]) fuse the features of all views $\{\mathbf{Z}^m\}_{m=1}^M$ to obtain a common representation for all views. In this way, the multi-view clustering task is transformed to single-view clustering task by conducting clustering directly on the fused features. However, the features of each view $\mathbf{Z}^m$ contain the common semantics as well as the view-private information. The latter is meaningless or even misleading, which might interfere with the quality of fused features and result in poor clustering effectiveness. (2) Some MVC methods (*e.g*., [8, 21]) learn consistent multi-view features to explore the common semantics by conducting a consistency objective on $\{\mathbf{Z}^m\}_{m=1}^M$, *e.g*., minimizing the distance of correlational features across all views. However, they also apply Eq. (1) to punish constraints on $\{\mathbf{Z}^m\}_{m=1}^M$ to avoid the model collapse and producing trivial solutions [11, 21]. The consistency objective and the reconstruction objective are pushed on the same features, so that their conflict may limit the quality of $\{\mathbf{Z}^m\}_{m=1}^M$. For example, the consistency objective aims to learn the common semantics while the reconstruction objective hopes to maintain the view-private information.

Recently, contrastive learning becomes popular and can be applied to achieve the consistency objective for multiple views. For instance, Trosten *et al*. [39] proposed a one-stage contrastive MVC method but its feature fusion suffers from challenge (1). Lin *et al*. [21] presented a two-stage contrastive MVC method by learning consistent features, but it does not consider challenge (2). Additionally, many contrastive learning methods (*e.g*., [19, 30, 40]) mainly handle single-view data with data augmentation. Such specific structure makes it difficult be applied in multi-view scenarios.

To address the aforementioned challenges, we propose a new framework of multi-level feature learning for contrastive multi-view clustering (named MFLVC) as shown in Figure 1. Specially, to reduce the adverse influence of view-private information, our framework avoids the direct feature fusion and builds a multi-level feature learning model for each view. To alleviate the conflict between the consistency objective and the reconstruction objective, we propose to conduct them in different feature spaces, where the consistency objective is achieved by the following multi-view contrastive learning.

3.2. Multi-view Contrastive Learning

Since the features $\{\mathbf{Z}^m\}_{m=1}^M$ obtained by Eq. (1) mix the common semantics with the view-private information, we treat $\{\mathbf{Z}^m\}_{m=1}^M$ as low-level features and learn another

level of features, *i.e.*, high-level features. To do this, we stack a feature MLP on $\{\mathbf{Z}^m\}_{m=1}^M$ to obtain the high-level features $\{\mathbf{H}^m\}_{m=1}^M$, where $\mathbf{h}_i^m \in \mathbb{R}^H$ and the feature MLP is a one-layer linear MLP denoted by $F(\{\mathbf{Z}^m\}_{m=1}^M; \mathbf{W}_H)$. In the low-level feature space, we leverage the reconstruction objective Eq. (1) to preserve the representation ability of $\{\mathbf{Z}^m\}_{m=1}^M$ so as to avoid the issue of model collapse. In the high-level feature space, we further achieve the consistency objective by contrastive learning to make $\{\mathbf{H}^m\}_{m=1}^M$ focus on learning the common semantics across all views.

Specifically, each high-level feature $\mathbf{h}_i^m$ has $(MN-1)$ feature pairs, *i.e.*, $\{\mathbf{h}_i^m, \mathbf{h}_j^n\}_{j=1,\dots,N}^{n=1,\dots,M}$, where $\{\mathbf{h}_i^m, \mathbf{h}_i^n\}_{n\neq m}$ are $(M-1)$ positive feature pairs and the rest $M(N-1)$ feature pairs are negative feature pairs. In contrastive learning, the similarities of positive pairs should be maximized and that of negative pairs should be minimized. Inspired by NT-Xent [7], the cosine distance is applied to measure the similarity between two features:

$$d(\mathbf{h}_i^m, \mathbf{h}_j^n) = \frac{\langle \mathbf{h}_i^m, \mathbf{h}_j^n \rangle}{\|\mathbf{h}_i^m\| \|\mathbf{h}_j^n\|}, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is dot product operator. Then, the feature contrastive loss between $\mathbf{H}^m$ and $\mathbf{H}^n$ is formulated as:

$$\ell_{fc}^{(mn)} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{d(\mathbf{h}_i^m, \mathbf{h}_i^n)/\tau_F}}{\sum_{j=1}^{N}\sum_{v=m,n} e^{d(\mathbf{h}_i^m, \mathbf{h}_j^v)/\tau_F} - e^{1/\tau_F}}, \tag{3}$$

where τ_F denotes the temperature parameter. In this paper, we design an accumulated multi-view feature contrastive loss across all views as:

$$\mathcal{L}_{\mathbf{H}} = \frac{1}{2}\sum_{m=1}^{M}\sum_{n\neq m} \ell_{fc}^{(mn)}. \tag{4}$$

In consequence, the features of each view can be written as $\mathbf{H}^m = \mathbf{W}_H \mathbf{Z}^m = \mathbf{W}_H E^m(\mathbf{X}^m)$. The encoder E^m is conducive to filtering out the random noise of $\mathbf{X}^m$. The reconstruction objective on $\mathbf{Z}^m$ avoids the model collapse as well as pushes both the common semantics and view-private information to be preserved in $\mathbf{Z}^m$. $\mathbf{W}_H$ is conducive to filtering out the view-private information of $\{\mathbf{Z}^m\}_{m=1}^M$. The consistency objective on $\{\mathbf{H}^m\}_{m=1}^M$ allows them to mine the common semantics across all views. As a result, the clusters of high-level features are close to the true semantic clusters. Intuitively, semantic information is a high-level concept that does not involve meaningless noise. Therefore, the high-level features within the same cluster are close to each other, resulting in dense shapes (verified in Sec. 5.1).

Learning semantic labels. This part explains how to obtain semantic labels for end-to-end clustering from the raw features in a fusion-free model. Specifically, we obtain the cluster assignments for all views $\{\mathbf{Q}^m \in \mathbb{R}^{N\times K}\}_{m=1}^M$ via a shared label MLP stacked on the low-level features, *i.e.*, $L(\{\mathbf{Z}^m\}_{m=1}^M; \mathbf{W}_Q)$. The last layer of the label MLP is set to the Softmax operation to output the probability, *e.g.*, q_{ij}^m represents the probability that the i-th sample belongs to the j-th cluster in the m-th view. Hence, the semantic label is identified by the largest element in a cluster assignment.

In real-world scenarios, however, some views of a sample might have wrong cluster labels due to the misleading of view-private information. In order to obtain robustness, we need to achieve clustering consistency, *i.e.*, the same cluster labels of all views represent the same semantic cluster. In other words, $\{\mathbf{Q}_{\cdot j}^m\}_{m=1}^M$ ($\mathbf{Q}_{\cdot j}^m \in \mathbb{R}^N$) need to be consistent. Similar to learning the high-level features, we adopt contrastive learning to achieve this consistency objective. For the m-th view, the same cluster labels $\mathbf{Q}_{\cdot j}^m$ have $(MK-1)$ label pairs, *i.e.*, $\{\mathbf{Q}_{\cdot j}^m, \mathbf{Q}_{\cdot k}^n\}_{k=1,\dots,K}^{n=1,\dots,M}$, where $\{\mathbf{Q}_{\cdot j}^m, \mathbf{Q}_{\cdot j}^n\}_{n\neq m}$ are constructed as $(M-1)$ positive label pairs and the rest $M(K-1)$ label pairs are negative label pairs. We further define the label contrastive loss between $\mathbf{Q}^m$ and $\mathbf{Q}^n$ as:

$$\ell_{lc}^{(mn)} = -\frac{1}{K}\sum_{j=1}^{K} \log \frac{e^{d(\mathbf{Q}_{\cdot j}^m, \mathbf{Q}_{\cdot j}^n)/\tau_L}}{\sum_{k=1}^{K}\sum_{v=m,n} e^{d(\mathbf{Q}_{\cdot j}^m, \mathbf{Q}_{\cdot k}^v)/\tau_L} - e^{1/\tau_L}}, \tag{5}$$

where τ_L denotes the temperature parameter. As thus, the clustering-oriented consistency objective is defined by:

$$\mathcal{L}_{\mathbf{Q}} = \frac{1}{2}\sum_{m=1}^{M}\sum_{n\neq m} \ell_{lc}^{(mn)} + \sum_{m=1}^{M}\sum_{j=1}^{K} s_j^m \log s_j^m, \tag{6}$$

where $s_j^m = \frac{1}{N}\sum_{i=1}^{N} q_{ij}^m$. The first part of Eq. (6) aims to learn the clustering consistency for all views. The second part of Eq. (6) is a regularization term [40], which is usually used to avoid all samples being assigned into a single cluster.

Overall, the loss of our multi-view contrastive learning consists of three parts:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\mathbf{Z}} + \mathcal{L}_{\mathbf{H}} + \mathcal{L}_{\mathbf{Q}} \\ &= \mathcal{L}_{\mathbf{Z}}(\{\mathbf{X}^m, \hat{\mathbf{X}}^m\}_{m=1}^M; \{\theta^m, \phi^m\}_{m=1}^M) \\ &+ \mathcal{L}_{\mathbf{H}}(\{\mathbf{H}^m\}_{m=1}^M; \mathbf{W}_H, \{\theta^m\}_{m=1}^M) \\ &+ \mathcal{L}_{\mathbf{Q}}(\{\mathbf{Q}^m\}_{m=1}^M; \mathbf{W}_Q, \{\theta^m\}_{m=1}^M), \end{aligned} \tag{7}$$

where $\mathcal{L}_{\mathbf{Z}}$ is the reconstruction objective conducted on the low-level features $\{\mathbf{Z}^m\}_{m=1}^M$ to avoid the model collapse. The consistency objectives $\mathcal{L}_{\mathbf{H}}$ and $\mathcal{L}_{\mathbf{Q}}$ are designed to learn the high-level features and the cluster assignments, respectively. We learn $\{\mathbf{Q}^m\}_{m=1}^M$ from $\{\mathbf{Z}^m\}_{m=1}^M$ rather than from $\{\mathbf{H}^m\}_{m=1}^M$ as it can avoid the influence between $\mathbf{W}_H$ and $\mathbf{W}_Q$. Meanwhile, $\mathbf{W}_H$ and $\mathbf{W}_Q$ will not be influenced by the gradient of $\mathcal{L}_{\mathbf{Z}}$. Thanks to this multi-level feature learning structure, we do not need weight parameters to balance the different losses in Eq. (7) (verified in Sec. 5.1).

3.3. Semantic Clustering with High-level Features

Through the multi-view contrastive learning, the model simultaneously learns the high-level features $\{\mathbf{H}^m\}_{m=1}^M$ and

the consistent cluster assignments $\{\mathbf{Q}^m\}_{m=1}^M$. We then treat $\{\mathbf{Q}^m\}_{m=1}^M$ as anchors and match them with the clusters among $\{\mathbf{H}^m\}_{m=1}^M$. In this way, we can leverage the cluster information contained in the high-level features to improve the clustering effectiveness of the semantic labels.

Concretely, we adopt K-means [26] to obtain the cluster information of each view. For the m-th view, letting $\{\mathbf{c}_k^m\}_{k=1}^K \in \mathbb{R}^H$ denote the K cluster centroids, we have:

$$\min_{\mathbf{c}_1^m,\mathbf{c}_2^m,\ldots,\mathbf{c}_K^m} \sum_{i=1}^{N}\sum_{j=1}^{K} \left\|\mathbf{h}_i^m - \mathbf{c}_j^m\right\|_2^2. \tag{8}$$

The cluster labels of all samples $\mathbf{p}^m \in \mathbb{R}^N$ are obtained by:

$$p_i^m = \operatorname*{argmin}_j \left\|\mathbf{h}_i^m - \mathbf{c}_j^m\right\|_2^2. \tag{9}$$

Let $\mathbf{l}^m \in \mathbb{R}^N$ denote the cluster labels outputted by the label MLP, where $l_i^m = \operatorname{argmax}_j q_{ij}^m$, it is worth noting that the clusters represented by $\mathbf{p}^m$ and $\mathbf{l}^m$ are not corresponding to each other. Because the clustering consistency is achieved by Eq. (6) in advance, l_i^m and l_i^n represent the same cluster. Therefore, we can treat $\mathbf{l}^m$ as anchors to modify $\mathbf{p}^m$ by the following maximum matching formula:

$$\begin{aligned} &\min_{\mathbf{A}^m} \mathbf{M}^m \mathbf{A}^m, \\ s.t.\ &\sum_{i=1} a_{ij}^m = 1, \sum_{j=1} a_{ij}^m = 1, \\ &a_{ij}^m \in \{0,1\}, i,j = 1,2,...,K, \end{aligned} \tag{10}$$

where $\mathbf{A}^m \in \{0,1\}^{K\times K}$ is the boolean matrix and $\mathbf{M}^m \in \mathbb{R}^{K\times K}$ denotes the cost matrix. $\mathbf{M}^m = \max_{i,j} \tilde{m}_{ij}^m - \tilde{\mathbf{M}}^m$ and $\tilde{m}_{ij}^m = \sum_{n=1}^N \mathbb{1}[l_n^m = i]\mathbb{1}[p_n^m = j]$, where $\mathbb{1}[\cdot]$ represents the indicator function. Eq. (10) can be optimized by the Hungarian algorithm [16]. The modified cluster assignments $\hat{\mathbf{p}}_i^m \in \{0,1\}^K$ for the i-th sample is defined as a one-hot vector. The k-th element of $\hat{\mathbf{p}}_i^m$ is 1 when k satisfies $k = k\mathbb{1}[a_{ks}^m = 1]\mathbb{1}[p_i^m = s], k, s \in \{1,2,\ldots,K\}$. We then fine-tune the model by cross-entropy loss:

$$\mathcal{L}_{\mathbf{P}} = -\sum_{m=1}^{M} \hat{\mathbf{P}}^m \log \mathbf{Q}^m, \tag{11}$$

where $\hat{\mathbf{P}}^m = [\hat{\mathbf{p}}_1^m; \hat{\mathbf{p}}_2^m; \ldots; \hat{\mathbf{p}}_N^m] \in \mathbb{R}^{N\times K}$. In this way, we can transfer the learned semantic knowledge to improve the clustering. Finally, the semantic label of the i-th sample is:

$$y_i = \operatorname*{argmax}_j \left(\frac{1}{M}\sum_{m=1}^{M} q_{ij}^m\right). \tag{12}$$

Optimization. The full optimization process of MFLVC is summarized in Algorithm 1. To be specific, we adopt the algorithm of mini-batch gradient descent to train the model, which consists of multiple autoencoders, a feature MLP, and a label MLP. The autoencoders are initialized by Eq. (1).

Algorithm 1 : The optimization of MFLVC

Input: Multi-view dataset $\{\mathbf{X}^m\}_{m=1}^M$; Number of clusters K; Temperature parameters τ_F and τ_L.

1: Initialize $\{\theta^m, \phi^m\}_{m=1}^M$ by minimizing Eq. (1).
2: Optimize $\mathbf{W}_H$, $\mathbf{W}_Q$, $\{\theta^m, \phi^m\}_{m=1}^M$ by Eq. (7).
3: Compute cluster labels by Eqs. (8) and (9).
4: Match multi-view cluster labels by solving Eq. (10).
5: Fine-tune $\mathbf{W}_Q$, $\{\theta^m\}_{m=1}^M$ by minimizing Eq. (11).
6: Calculate semantic labels by Eq. (12).

Output: The label predictor $\{\{\theta^m\}_{m=1}^M, \mathbf{W}_Q\}$; The high-level feature extractor $\{\{\theta^m\}_{m=1}^M, \mathbf{W}_H\}$.

The multi-view contrastive learning is then conducted to achieve the common semantics and clustering consistency by Eq. (7). After performing the multi-view contrastive learning, the cluster labels obtained from high-level features are modified through the maximum matching formula in Eq. (10). The modified cluster labels are then used to fine-tune the model by Eq. (11). The high-level feature extractor includes the encoders and the feature MLP, while the label predictor includes the encoders and the label MLP.

4. Experiments

4.1. Experimental Setup

Datasets	#Samples	#Views	#Classes
MNIST-USPS	5,000	2	10
BDGP	2,500	2	5
CCV	6,773	3	20
Fashion	10,000	3	10
Caltech-2V	1,400	2	7
Caltech-3V	1,400	3	7
Caltech-4V	1,400	4	7
Caltech-5V	1,400	5	7

Table 1. The information of the datasets in our experiments.

Datasets. The experiments are carried out on the five public datasets as shown in Table 1. *MNIST-USPS* [34] is a popular handwritten digit dataset, which contains 5,000 samples with two different styles of digital images. *BDGP* [5] contains 2,500 samples of drosophila embryos, each of which is represented by visual and textual features. *Columbia Consumer Video (CCV)* [15] is a video dataset with 6,773 samples belonging to 20 classes and provides hand-crafted Bag-of-Words representations of three views, such as STIP, SIFT, and MFCC. *Fashion* [47] is an image dataset about products, where we follow the literature [50] to treat different three styles as three views of one product. *Caltech* [9] is a RGB image dataset with multiple views, based on which we build four datasets for evaluating the robustness of the comparison methods in terms of the number of views. Concretely,

Datasets	MNIST-USPS			BDGP			CCV			Fashion		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
RMSL [18] (2019)	0.424	0.318	0.428	0.849	0.630	0.849	0.215	0.157	0.243	0.408	0.405	0.421
MVC-LFA [41] (2019)	0.768	0.675	0.768	0.564	0.395	0.612	0.232	0.195	0.261	0.791	0.759	0.794
COMIC [34] (2019)	0.482	0.709	0.531	0.578	0.642	0.639	0.157	0.081	0.157	0.578	0.642	0.608
CDIMC-net [44] (2020)	0.620	0.676	0.647	0.884	0.799	0.885	0.201	0.171	0.218	0.776	0.809	0.789
EAMC [59] (2020)	0.735	0.837	0.778	0.681	0.480	0.697	0.263	0.267	0.274	0.614	0.608	0.638
IMVTSC-MVI [46] (2021)	0.669	0.592	0.717	0.981	0.950	0.982	0.117	0.060	0.158	0.632	0.648	0.635
SiMVC [39] (2021)	0.981	0.962	0.981	0.704	0.545	0.723	0.151	0.125	0.216	0.825	0.839	0.825
CoMVC [39] (2021)	0.987	0.976	0.989	0.802	0.670	0.803	0.296	0.286	0.297	0.857	0.864	0.863
MFLVC (ours)	**0.995**	**0.985**	**0.995**	**0.989**	**0.966**	**0.989**	**0.312**	**0.316**	**0.339**	**0.992**	**0.980**	**0.992**

Table 2. Results of all methods on four datasets. Bold denotes the best results and underline denotes the second-best.

Datasets	Caltech-2V			Caltech-3V			Caltech-4V			Caltech-5V		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
RMSL [18] (2019)	0.525	0.474	0.540	0.554	0.480	0.554	0.596	0.551	0.608	0.354	0.340	0.391
MVC-LFA [41] (2019)	0.462	0.348	0.496	0.551	0.423	0.578	0.609	0.522	0.636	0.741	0.601	0.747
COMIC [34] (2019)	0.422	0.446	0.535	0.447	0.491	0.575	0.637	0.609	**0.764**	0.532	0.549	0.604
CDIMC-net [44] (2020)	0.515	0.480	0.564	0.528	0.483	0.565	0.560	0.564	0.617	0.727	0.692	0.742
EAMC [59] (2020)	0.419	0.256	0.427	0.389	0.214	0.398	0.356	0.205	0.370	0.318	0.173	0.342
IMVTSC-MVI [46] (2021)	0.490	0.398	0.540	0.558	0.445	0.576	0.687	0.610	0.719	0.760	0.691	0.785
SiMVC [39] (2021)	0.508	0.471	0.557	0.569	0.495	0.591	0.619	0.536	0.630	0.719	0.677	0.729
CoMVC [39] (2021)	0.466	0.426	0.527	0.541	0.504	0.584	0.568	0.569	0.646	0.700	0.687	0.746
MFLVC (ours)	**0.606**	**0.528**	**0.616**	**0.631**	**0.566**	**0.639**	**0.733**	**0.652**	0.734	**0.804**	**0.703**	**0.804**

Table 3. Results of all methods on Caltech with different views. "-XV" represents that there are X views.

Caltech-2V includes WM and CENTRIST; *Caltech-3V* includes WM, CENTRIST, and LBP; *Caltech-4V* includes WM, CENTRIST, LBP, and GIST; *Caltech-5V* includes WM, CENTRIST, LBP, GIST, and HOG.

Implementation. All the datasets are reshaped into vectors, and the fully connected networks with the similar architecture are adopted to implement the autoencoders for all views in our MFLVC. Adam optimizer [17] is adopted for optimization. The code of MFLVC is implemented by PyTorch [32]. More implementation details are provided in https://github.com/SubmissionsIn/MFLVC.

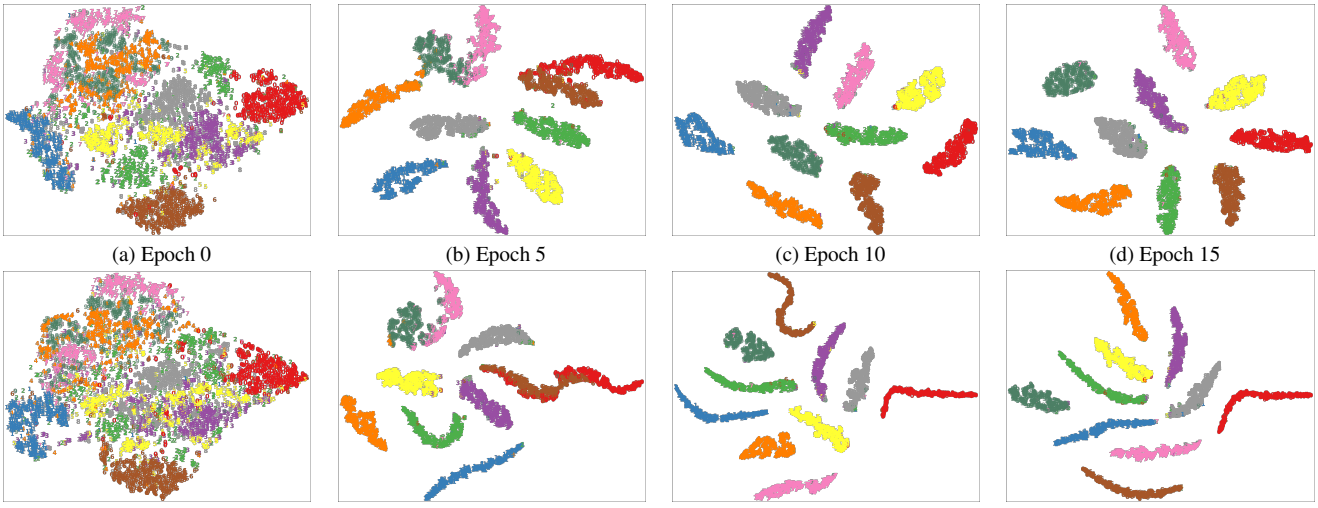
Comparison methods. The comparison methods include classical and state-of-the-art methods, *i.e.*, 4 traditional methods (RMSL [18], MVC-LFA [41], COMIC [34], and IMVTSC-MVI [46]) and 4 deep methods (CDIMC-net [44], EAMC [59], SiMVC [39], and CoMVC [39]).

Evaluation metrics. The clustering effectiveness is evaluated by three metrics, *i.e.*, clustering accuracy (ACC), normalized mutual information (NMI), and purity (PUR). The mean values of 10 runs are reported for all methods.

4.2. Result Analysis

The comparison results on four datasets are shown in Table 2, where many comparison methods (*e.g.*, RMSL and COMIC) punish multiple objectives on the same features, and CDIMC-net, EAMC, SiMVC, and CoMVC are feature fusion methods. One could find that: (1) Our MFLVC achieves the best performance in terms of all metrics. Especially on Dataset Fashion, MFLVC outperforms the best comparison method CoMVC (*i.e.*, 85%) by about 14% in terms of ACC. This is because our model is fusion-free and it conducts the reconstruction objective and the consistency objective in different feature spaces so that the adverse influence of view-private information can be reduced. (2) The improvements obtained by the previous contrastive MVC method (*i.e.*, CoMVC) are limited. Our MFLVC is also a contrastive MVC method, instead, it avoids the fusion of view-private information and its multi-level feature learning framework allows the high-level features to learn the common semantics across all views more effectively.

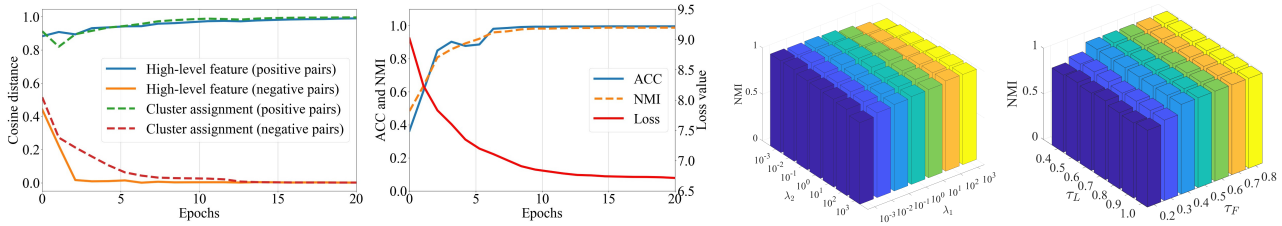
To further verify our method, we build four datasets based on Caltech and test the performance of all comparison methods. Table 3 shows the results on Caltech with different views, from which we could have the following observations: (1) The clustering effectiveness of most methods improves with the increase of the number of views, *i.e.*, ACC increases from 60% to 80%. (2) Compared to 8 comparison methods, our MFLVC mostly achieves the best performance indicating its robustness. (3) Some methods obtain bad results when increasing the number of views. For example, RMSL, COMIC, and EAMC achieve ACC about 35%, 53%, and 31% on Caltech-5V which are lower than that on Caltech-4V



(a) Epoch 0 (b) Epoch 5 (c) Epoch 10 (d) Epoch 15

(e) Epoch 0 (f) Epoch 5 (g) Epoch 10 (h) Epoch 15

Figure 2. Visualization of low-level features (a-d) and high-level features (e-h) for the contrastive learning process.



(a) Similarities $vs.$ contrastive learning (b) Loss $vs.$ performance (c) λ_1 $vs.$ λ_2 (d) τ_F $vs.$ τ_L

Figure 3. (a) The similarities of feature pairs and label pairs. (b) Convergence analysis. (c) and (d) Parameters sensitivity analysis.

(*i.e.*, 59%, 63%, and 35%). The reason is that the data of each view simultaneously contain useful common semantics as well as meaningless view-private information. Views contain much view-private information which might increase the difficulty of extracting their common semantics. These observations further verify the effectiveness of our method, which learns multiple levels of features so as to reduce the interference from the view-private information.

5. Model Analysis

5.1. Understand the Multi-level Feature Learning

In order to investigate the proposed multi-level feature learning, we take MNIST-USPS as an example and visualize its training process. The MNIST view is shown in Figure 2 via t-SNE [25]. It can be discovered that the cluster structures of low-level features and high-level features become clear during the training process. The clusters of low-level features are not dense. This is because the low-level features have maintained the diversity among samples by reconstruction objective. In contrast, the clusters of high-level features are dense and have better low-dimensional manifolds. Additionally, in Figure 3(a), the similarities of positive feature pairs are rising while that of negative feature pairs are decreasing. This indicates that the information learned by the high-level features is close to the common semantics across multiple views. These observations are in agreement with our motivations, *i.e.*, the feature MLP can filter out the view-private information of multiple views so the outputted high-level features are in dense shapes. The similarities of positive label pairs are also rising which indicates that the clustering consistency of semantic labels is achieved.

Convergence analysis. It is not difficult to discover that the objectives of $\mathcal{L}_{\mathbf{Z}}$, $\mathcal{L}_{\mathbf{H}}$, $\mathcal{L}_{\mathbf{Q}}$, and $\mathcal{L}_{\mathbf{P}}$, *i.e.*, Eqs. (1,4,6,11) are all convex functions. As shown in Figure 3(b), the clustering effectiveness increases with the decrease of loss values, indicating that MFLVC enjoys good convergence property.

Parameter sensitivity analysis. We investigate whether hyper-parameters are needed to balance the loss components in Eq. (7), *i.e.*, $\mathcal{L}_{\mathbf{Z}} + \lambda_1\mathcal{L}_{\mathbf{H}} + \lambda_2\mathcal{L}_{\mathbf{Q}}$. Figure 3(c) shows the mean values of NMI within 10 independent runs, which indicates that our model is insensitive to λ_1 and λ_2. This is because our model has a well-designed multi-level feature learning framework, by which the interference among differ-

	Components			MNIST-USPS		BDGP	
	$\mathcal{L}_{\mathbf{Q}}$	$\mathcal{L}_{\mathbf{Z}}$	$\mathcal{L}_{\mathbf{H}}$ and $\mathcal{L}_{\mathbf{P}}$	ACC	NMI	ACC	NMI
(A)	✓			0.676	0.777	0.715	0.663
(B)	✓	✓		0.891	0.939	0.825	0.690
(C)	✓		✓	0.984	0.962	0.955	0.886
(D)	✓	✓	✓	0.995	0.985	0.989	0.966

Table 4. Ablation studies on loss components.

		MNIST-USPS		BDGP	
		ACC	NMI	ACC	NMI
(a)	$\mathbf{X} - \mathbf{Q}_{\checkmark}$	0.676	0.777	0.715	0.663
(b)	$\mathbf{X} - \mathbf{Z}_{\checkmark} - \mathbf{Q}_{\checkmark}$	0.921	0.860	0.652	0.498
(c)	$\mathbf{X} - \mathbf{Z}_{\checkmark} - \mathbf{H}_{\checkmark} - \mathbf{Q}_{\checkmark}$	0.948	0.894	0.742	0.654
(d)	$\mathbf{X} - \mathbf{Z}_{\times} - \mathbf{H}_{\checkmark} - \mathbf{Q}_{\checkmark}$	0.995	0.985	0.989	0.966

Table 5. Ablation studies on contrastive learning structures. "✓" represents that the contrastive loss is optimized on the features.

ent features can be reduced. In this paper, we set $\lambda_1 = 1.0$ and $\lambda_2 = 1.0$ for all used datasets. Furthermore, the multi-view contrastive learning includes two temperature parameters, *i.e.*, τ_F of the feature contrastive loss in Eq. (3) and τ_L of the label contrastive loss in Eq. (5). Figure 3(d) indicates that our model is insensitive to the choice of τ_F and τ_L. Empirically, we set $\tau_F = 0.5$ and $\tau_L = 1.0$.

5.2. Ablation Studies

Loss components. We conduct ablation studies on the loss components in Eq. (7) and Eq. (11) to investigate their effectiveness. Table 4 shows different loss components and the corresponding experimental results. (A) $\mathcal{L}_{\mathbf{Q}}$ is optimized to achieve the basic goal of multi-view clustering, *i.e.*, learning the clustering consistency. (B) $\mathcal{L}_{\mathbf{Z}}$ is optimized to make the low-level features be capable of reconstructing the multiple views. (C) $\mathcal{L}_{\mathbf{H}}$ is optimized to learn the high-level features, which are then used to fine-tune the semantic labels by $\mathcal{L}_{\mathbf{P}}$. (D) The complete loss components of our method. In terms of the results, (B) and (D) have better performance than (A) and (C), respectively, indicating that the reconstruction objective is important. Especially when the model has only low-level features, the results of (B) are better than that of (A) by about 20% and 10% on MNIST-USPS and BDGP, respectively. According to (C) and (D), we can find that the learned high-level features play the most important role in improving the clustering effectiveness. For example, the results of (C) are better than that of (A) by about 30% and 20% on MNIST-USPS and BDGP, respectively.

Contrastive learning structures. To further verify our proposal, we perform contrastive learning (*i.e.*, consistency objective) on different network structures. As shown in Table 5, (a) The semantic labels $\mathbf{Q}$ are learned directly from the input features $\mathbf{X}$. This structure is similar to [29, 40, 58] in some degree. It results in poor performance by directly extending contrastive learning to the multi-view scenarios. (b) Between $\mathbf{X}$ and $\mathbf{Q}$, we set the low-level features $\mathbf{Z}$ and perform contrastive learning on $\mathbf{Q}$ and $\mathbf{Z}$. This structure is similar to [19, 21, 39] in some degree and the performance is also limited. (c) Based on $\mathbf{Z}$, we stack a feature MLP to obtain the high-level features $\mathbf{H}$ and perform contrastive learning on $\mathbf{Z}$, $\mathbf{H}$, and $\mathbf{Q}$. As for (b) and (c), the reconstruction objective is also performed on $\mathbf{Z}$. (b) and (c) make progress on MNIST-USPS, because the two views of MNIST-USPS are digital images and they have little view-private information to influence the learning performance. However, (b) and (c) cannot mine the common semantics well on BDGP. The reason is that the two views of BDGP are visual features and text features and they have much view-private information. It results in poor performance when performing reconstruction and consistency objectives on the same features (*i.e.*, $\mathbf{Z}$). (d) We perform contrastive learning only on $\mathbf{H}$ and $\mathbf{Q}$ while leaving reconstruction objective on $\mathbf{Z}$. This setting obtains the best performance by performing consistency and reconstruction objectives in different feature spaces. These experiments further verified the effectiveness of our method, and confirmed that it is useful to learn representations via a multi-level feature learning structure.

6. Conclusion

In this paper, we have proposed a new framework of multi-level feature learning for contrastive multi-view clustering. For each view, the proposed framework learns multiple levels of features, including low-level features, high-level features, and semantic labels in a fusion-free manner. This allows our model to learn the common semantics across all views and reduce the adverse influence of view-private information. Extensive experiments on five public datasets demonstrate that our method obtains state-of-the-art performance.

Broader impacts. The proposed framework learned a high-level feature extractor and a label predictor, which can be applied to downstream tasks such as feature compression, unsupervised labeling, and cross-modal retrieval, *etc*. However, this work aims to provide a general framework and the trained model might be affected by the intrinsic bias of data especially with dirty samples. Therefore, the future works could extend our framework to other application scenarios.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grants No. 61806043 and No. 61876046) and the Guangxi "Bagui" Teams for Innovation and Research, China. Lifang He was supported by the Lehigh's Accelerator Grant (No. S00010293).

References

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

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, pages 9758–9770, 2019. 1

[3] Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, pages 4660–4671, 2020. 1

[4] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *IJCAI*, pages 2598–2604, 2013. 3

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

[6] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *CVPR*, pages 586–594, 2015. 1, 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 3, 4

[8] Jiafeng Cheng, Qianqian Wang, Zhiqiang Tao, De-Yan Xie, and Quanxue Gao. Multi-view attribute graph convolution networks for clustering. In *IJCAI*, pages 2973–2979, 2020. 3

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR*, pages 178–178, 2004. 5

[10] Jun Guo and Jiahui Ye. Anchors bring ease: An embarrassingly simple approach to partial multi-view clustering. In *AAAI*, pages 118–125, 2019. 1

[11] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *IJCAI*, pages 1753–1759, 2017. 3

[12] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *ICML*, pages 4116–4126, 2020. 3

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

[14] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, pages 9248–9257, 2019. 1

[15] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ICMR*, pages 1–8, 2011. 5

[16] Roy Jonker and Ton Volgenant. Improving the hungarian assignment algorithm. *Operations Research Letters*, 5(4):171–175, 1986. 5

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Ruihuang Li, Changqing Zhang, Huazhu Fu, Xi Peng, Tianyi Zhou, and Qinghua Hu. Reciprocal multi-layer subspace learning for multi-view clustering. In *ICCV*, pages 8172–8180, 2019. 1, 2, 6

[19] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *AAAI*, pages 8547–8555, 2021. 3, 8

[20] Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, and Zhaohua Yang. Deep adversarial multi-view clustering network. In *IJCAI*, pages 2952–2958, 2019. 1, 3

[21] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. COMPLETER: Incomplete multi-view clustering via contrastive prediction. In *CVPR*, 2021. 1, 3, 8

[22] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Li Liu, Siqi Wang, Weixuan Liang, and Jiangyong Shi. One-pass multi-view clustering for large-scale data. In *ICCV*, pages 12344–12353, 2021. 1

[23] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. Multi-view clustering via joint nonnegative matrix factorization. In *SDM*, pages 252–260, 2013. 2

[24] Shirui Luo, Changqing Zhang, Wei Zhang, and Xiaochun Cao. Consistent and specific multi-view subspace clustering. In *AAAI*, 2018. 1, 2

[25] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 7

[26] James MacQueen. Some methods for classification and analysis of multivariate observations. In *BSMSP*, pages 281–297, 1967. 3, 5

[27] Kevis-Kokitsi Maninis, Stefan Popov, Matthias Niesser, and Vittorio Ferrari. Vid2cad: Cad model alignment using multi-view constraints from videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[28] Feiping Nie, Jing Li, and Xuelong Li. Self-weighted multi-view clustering with multiple graphs. In *IJCAI*, pages 2564–2570, 2017. 1, 3

[29] Chuang Niu and Ge Wang. SPICE: Semantic pseudo-labeling for image clustering. *arXiv preprint arXiv:2103.09382*, 2021. 3, 8

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[31] Erlin Pan and Zhao Kang. Multi-view contrastive graph clustering. In *NeurIPS*, 2021. 3

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 6

[33] Liang Peng, Yang Yang, Zheng Wang, Zi Huang, and Heng Tao Shen. MRA-Net: Improving vqa via multi-modal relation attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):318–329, 2020. 1

[34] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. COMIC: Multi-view clustering without parameter selection. In *ICML*, pages 5092–5101, 2019. 3, 5, 6

[35] Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Nataša Sladoje. CoMIR: Contrastive multimodal image representation for registration. In *NeurIPS*, pages 18433–18444, 2020. 3

[36] Raeid Saqur and Karthik Narasimhan. Multimodal graph networks for compositional generalization in visual question answering. In *NeurIPS*, pages 3070–3081, 2020. 1

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

[38] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pages 776–794, 2020. 3

[39] Daniel J. Trosten, Sigurd Løkse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *CVPR*, pages 1255–1265, 2021. 1, 3, 6, 8

[40] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. SCAN: Learning to classify images without labels. In *ECCV*, pages 268–285, 2020. 3, 4, 8

[41] Siwei Wang, Xinwang Liu, En Zhu, Chang Tang, Jiyuan Liu, Jingtao Hu, Jingyuan Xia, and Jianping Yin. Multi-view clustering via late fusion alignment maximization. In *IJCAI*, pages 3778–3784, 2019. 6

[42] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pages 9929–9939, 2020. 3

[43] Jiwei Wei, Yang Yang, Xing Xu, Xiaofeng Zhu, and Heng Tao Shen. Universal weighting metric learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[44] Jie Wen, Zheng Zhang, Yong Xu, Bob Zhang, Lunke Fei, and Guo-Sen Xie. CDIMC-net: Cognitive deep incomplete multi-view clustering network. In *IJCAI*, pages 3230–3236, 2020. 1, 6

[45] Jie Wen, Zheng Zhang, Yong Xu, and Zuofeng Zhong. Incomplete multi-view clustering via graph regularized matrix factorization. In *ECCV Workshops*, 2018. 1

[46] Jie Wen, Zheng Zhang, Zhao Zhang, Lei Zhu, Lunke Fei, Bob Zhang, and Yong Xu. Unified tensor framework for incomplete multi-view clustering and missing-view inferring. In *AAAI*, pages 10273–10281, 2021. 6

[47] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 5

[48] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, pages 478–487, 2016. 1

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

[50] Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-VAE: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *ICCV*, pages 9234–9243, 2021. 1, 3, 5

[51] Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, and Xiaorong Pu. Self-supervised discriminative feature learning for deep multi-view clustering. *arXiv preprint arXiv:2103.15069*, 2021. 3

[52] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jian Cheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[53] Zuyuan Yang, Naiyao Liang, Wei Yan, Zhenni Li, and Shengli Xie. Uniform distribution non-negative matrix factorization for multiview clustering. *IEEE Transactions on Cybernetics*, pages 3249–3262, 2021. 1

[54] Ming Yin, Weitian Huang, and Junbin Gao. Shared generative latent representation learning for multi-view clustering. In *AAAI*, pages 6688–6695, 2020. 1

[55] Kun Zhan, Changqing Zhang, Junpeng Guan, and Junsheng Wang. Graph learning for multiview clustering. *IEEE Transactions on Cybernetics*, 48(10):2887–2895, 2017. 1

[56] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *AAAI*, pages 2921–2927, 2017. 1, 2

[57] Guo Zhong and Chi-Man Pun. Improved normalized cut for multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[58] Huasong Zhong, Chong Chen, Zhongming Jin, and Xian-Sheng Hua. Deep robust clustering by contrastive learning. *arXiv preprint arXiv:2008.03030*, 2020. 8

[59] Runwu Zhou and Yi-Dong Shen. End-to-end adversarial-attention network for multi-modal clustering. In *CVPR*, pages 14619–14628, 2020. 1, 3, 6

[60] Xiaofeng Zhu, Shichao Zhang, Wei He, Rongyao Hu, Cong Lei, and Pengfei Zhu. One-step multi-view spectral clustering. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):2022–2034, 2018. 1