Differentiable Information Bottleneck for Deterministic Multi-view Clustering

Xiaoqiang Yan, Zhixiang Jin*, Fengshou Han, Yangdong Ye*

School of Computer and Artificial Intelligence, Zhengzhou University

{iexqyan, ieydye}@zzu.edu.cn, {zxjin, iefshan}@gs.zzu.edu.cn

Abstract

In recent several years, the information bottleneck (IB) principle provides an information-theoretic framework for deep multi-view clustering (MVC) by compressing multi-view observations while preserving the relevant information of multiple views. Although existing IB-based deep MVC methods have achieved huge success, they rely on variational approximation and distribution assumption to estimate the lower bound of mutual information, which is a notoriously hard and impractical problem in high-dimensional multi-view spaces. In this work, we propose a new differentiable information bottleneck (DIB) method, which provides a deterministic and analytical MVC solution by fitting the mutual information without the necessity of variational approximation. Specifically, we first propose to directly fit the mutual information of high-dimensional spaces by leveraging normalized kernel Gram matrix, which does not require any auxiliary neural estimator to estimate the lower bound of mutual information. Then, based on the new mutual information measurement, a deterministic multi-view neural network with analytical gradients is explicitly trained to parameterize IB principle, which derives a deterministic compression of input variables from different views. Finally, a triplet consistency discovery mechanism is devised, which is capable of mining the feature consistency, cluster consistency and joint consistency based on the deterministic and compact representations. Extensive experimental results show the superiority of our DIB method on 6 benchmarks compared with 13 state-of-the-art baselines.

1. Introduction

Multi-view clustering (MVC) [8, 48] aims to discover hidden patterns or potential structures by leveraging the complementary information in multi-view data. In the literature, MVC involving traditional machine learning techniques can be classified into graph-based [14], subspace-based [18] and matrix factorization-based methods [22].

However, these traditional MVC based on shallow learning models often exhibits poor representation ability on large-scale high-dimensional and non-linear multi-view data. Recently, deep learning models have seen widespread adoption in MVC owing to their powerful representation capability, resulting in deep MVC [6, 15, 20, 21, 38, 43–45]. Although achieving promising performance, most existing deep MVC emphasizes the relevant correlations across multiple views and ignores the limitations of the irrelevant information in each view, such as noises, corruptions or even view-private attributes.

In addressing these challenges, several recent approaches have resorted to the notable information bottleneck (IB) principle to multi-view clustering [26, 41, 46]. By formulating mutual information (MI), IB provides an information-theoretic framework to learn a compact representation and remove irrelevant information for a given task [35]. Despite the successful applications, the estimation of mutual information is a notoriously hard problem in high-dimensional multi-view space since the complicated joint distribution of two variables is often criticized to be hard or impossible. To overcome this constraint, variational approximation offers a natural solution to construct and estimate a lower bound of the mutual information of high-dimensional variables [1, 40]. Inspired by this, IB and its variational versions have achieved promising performance by exploring the training dynamics in deep multi-view clustering and representation models as well as a learning objective [27, 28, 47]. However, the variational approximation in existing IB-based deep MVC results in the uncertainty in multi-view representation learning. Specifically, existing IB-based deep MVC leverages variational approximation to estimate the marginal and posterior probability distribution of the potential feature representations (as shown in Fig.1). In the process of variational approximation, IB-based deep MVC methods introduce an auxiliary neural network to estimate the mean and variance of the posterior distribution so as to fit the posterior distribution. Then, they impose $Kullback\text{-}Leibler$ (KL) divergence constraint between the posterior distribution and variational approximation to align them [1]. Thus, the approximation error introduced by vari-

---

* Corresponding author.

(a) Variational approximation
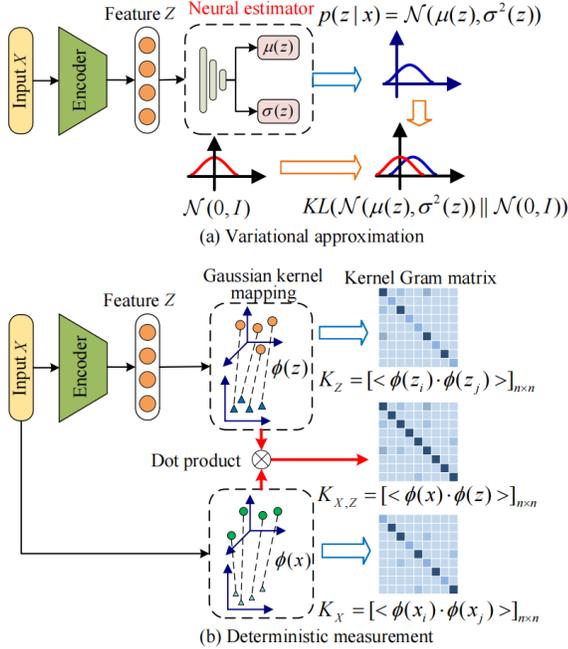
(b) Deterministic measurement

Figure 1. Variational approximation and deterministic measurement. (a) Variational approximation requires a neural estimator to estimate the posterior distribution $p(z|x)$ of the representation while assuming the marginal distribution of the representation follows a standard normal distribution, so as to approximate the lower bound of the mutual information. (b) Our deterministic measurement leverages the gaussian kernel function to construct the kernel Gram matrix which can measure the distance between data pairs. Then the eigenvalues of the Gram matrix can be expressed to entropy function (see section 3.2 for detailed proof).

ational approximation increases the uncertainty of mutual information estimation.

In this study, we propose a novel differentiable information bottleneck (DIB) method for deterministic and analytical multi-view clustering without variational approximation. As shown in Figure 2, DIB learns a latent and compact space for each view in a deterministic compression manner while capturing triplet consistency derived from high-level features and semantic labels across multiple views. To this end, we first design a novel MI measurement to directly fit the mutual information between high-dimensional multi-view spaces by leveraging normalized kernel Gram matrix, which can measure the information about feature representations directly from the original data and does not need any neural estimators to learn the lower bound of mutual information. Then, based on the proposed MI measurement without variational approximation, a deterministic multi-view neural network is explicitly trained to parameterize IB principle with analytical gradients, which derives a deterministic compression and learns a compact represen-

tation for each view. Finally, a unified objective function under our DIB framework is devised to optimize the deterministic compression and triplet consistency discovery simultaneously, in which consistent information of multiple views from high-level features and semantic labels is characterized based on the deterministic and compact representations. Extensive experimental results verify the effectiveness and promising performance of DIB compared with state-of-the-art baselines. In summary, this study makes the following contributions.

- We propose a novel differentiable information bottleneck (DIB) method for deterministic multi-view clustering, which provides a deterministic and analytical MVC solution by essentially fitting the mutual information without the necessity of variational approximation.
- A novel MI measurement without variational approximation is proposed to fit the mutual information of high-dimensional spaces directly by eigenvalues of the normalized kernel Gram matrix. This work is a valuable attempt to directly measure the information about feature representations from the data rather than building a neural estimator to approximate the lower bound of mutual information.
- A deterministic neural network with analytical gradients is built to parameterize IB principle, which enjoys a concise and tractable objective and provides a deterministic compression of input variables from different views.

2. Related Work and Preliminaries

2.1. Information Bottleneck

The information bottleneck (IB) [35] originates from rate-distortion and attempts to compress source variable $\mathbf{X}$ into its compressed representation $\mathbf{Z}$ while preserving the information that can predict relevant variable $\mathbf{Y}$. It is assumed that we have access to the joint distribution $p(\mathbf{X}, \mathbf{Y})$ with the goal of pursuing the following quantization

$$\max \mathbf{IB}_\beta = I(\mathbf{Z}; \mathbf{Y}) - \beta I(\mathbf{X}; \mathbf{Z}) \qquad (1)$$

where $I()$ is the mutual information and β is a trade-off parameter.

Recently, the idea of exploring a good representation with IB principle is becoming prevalent and it also achieves great success in deep multi-view learning, such as multi-view clustering [27, 28, 47, 49], multi-view representation learning [9, 37, 39], multi-view graph clustering [5]. However, both the "black box" operation of neural network and the approximate error introduced by variational approximation increase the uncertainty of mutual information estimation.

Different from the aforementioned approaches, DIB provides a deterministic and analytical MVC solution by essentially fitting the mutual information without the necessity

of variational approximation. Moreover, the new measurement of mutual information is differentiable and can explicitly parameterize IB principle with analytical gradients.

2.2. Deep Multi-view Clustering

Existing deep MVC approaches can be classified into the following categories: deep embedding-based, deep graph-based, deep adversarial-based and contrastive MVC. Deep embedding-based MVC jointly optimizes the embedded representation of multiple views and the clustering process [15, 42]. Deep graph-based MVC learns the cluster structures of multi-view data by forming a better graph from multiple views [21, 51]. Deep adversarial-based MVC uses adversarial training as a regularizer to align the multi-view data [20]. Contrastive MVC enables a better latent space of multiple views by characterizing the positive and negative samples [19, 43].

The proposed DIB is remarkably different from existing deep MVC approaches. First, DIB constructs a view-specific encoder with the constraint of differentiable mutual information, which can learn a compact and discriminative representation for each view by preserving relevant information and eliminating irrelevant information simultaneously. Second, DIB leverages a deterministic neural network with analytical gradients driven by the mutual information without variational approximation to parameterize IB principle, which enjoys a concise and tractable objective and provides a deterministic compression of input variables from different views. Third, a triplet consistency discovery mechanism under our DIB framework is devised, which capture the feature consistency, cluster consistency and joint consistency in a triplet manner.

3. Differentiable Information Bottleneck

3.1. Problem Statement

Problem statement. Given an unlabelled multi-view collection $\{\mathbf{X}^v \in \mathbb{R}^{N \times D^v}\}_{v=1}^V$, multi-view clustering aims to partition the data samples into K clusters, where V is the number of views, $x_i^v \in \mathbb{R}^{D^v}$ is the samples of the v-th view, N and D^v are the data size and feature dimension of the v-th view respectively.

Recently, the deep MVC approaches based on IB principle have achieved huge success since it provides an information-theoretic framework to learn a compact representation and remove irrelevant information for a given task. However, despite the successful applications, the approximate error introduced by variational approximation increases the uncertainty of mutual information estimation. Aiming at these issues, we propose a novel differentiable information bottleneck for deterministic MVC. Intuitively, DIB should meet the following requirements. 1) **Information measurement**. It should directly measure the in-

formation of original data about its feature representations and does not need any neural estimators to learn the lower bound of mutual information. 2) **Deterministic compression**. A neural network driven by the information measurement without variational approximation should have analytical gradients that allow us to parameterize the IB principle and optimize it through backward propagation. 3) **Consistency maximization**. DIB should characterize the consistency of multiple views more comprehensively based on the deterministic and compact representations. To facilitate these goals, we design the network architecture of DIB method as shown in Fig. 2. From this figure, we can see that DIB consists of deterministic compression and triplet consistency discovery. For convenience, we first provide the definition of the proposed DIB method.

Definition 1 (Differentiable information bottleneck, DIB). *Suppose there exists an unlabelled multi-view collection $\{\mathbf{X}^v \in \mathbb{R}^{N \times D^v}\}_{v=1}^V$, DIB consists of deterministic compression and triplet consistency discovery. The deterministic compression part learns a deterministic and compact representation $\{\mathbf{Z}^v\}_{v=1}^V$ for each view $\{\mathbf{X}^v\}_{v=1}^V$ using view-specific encoder E^v with an information-theoretic constraint. In the triplet consistency discovery part, we explore the consistency of multiple views from high-level features $\{\mathbf{H}^v\}_{v=1}^V$ and semantic labels $\{\mathbf{S}^v\}_{v=1}^V$ in a triplet manner. In summary, the goal of DIB is to search a reasonable clustering assignment $\mathbf{C}$ by learning a deterministic and compact representation for each view while maximally preserving the consistency across multiple views.*

3.2. Mutual Information without Variational Approximation

In this subsection, we design a novel mutual information measurement without variational approximation to directly fit the mutual information of high-dimensional spaces by leveraging normalized kernel Gram matrix. Specifically, we first show the eigenvalues of the kernel Gram matrix can be expressed by the recently proposed Rényi's α-order entropy [30]. Then, the mutual information can be achieved via matrix-based Rényi's α-order entropy function and the joint-entropy function without variational approximation. For convenience, we first provide the definition of Gram matrix.

Definition 2 (Gram matrix). *Given a set of vectors $\{v_i\}_{i=1}^n$ in an inner product space, the Gram matrix G is defined as an $n \times n$ matrix with entries*

$$G_{ij} = <v_i, v_j> \tag{2}$$

where $<v_i, v_j>$ denotes the inner product of vectors v_i and v_j.

Different from approximating the lower bound of mutual information through a neural estimator in variational
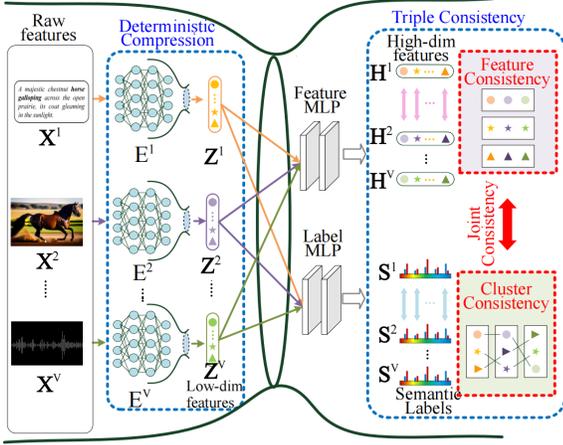
Figure 2. Framework of the DIB. In DIB, the deterministic compression aims to learn a compact representation for each view through the new mutual information measurement without variational approximation. The triplet consistency discovery mechanism is devised to mine the feature, cluster and joint consistency from the compact representation.

approximation [1], we utilize the eigenvalues of the kernel Gram matrix to directly fit the recently proposed Rényi's α-order entropy as shown in following proposition.

Proposition 1. *The Rényi's α-order entropy of a random variable $X \in \mathcal{R}^{N \times D^v}$ can be fitted by the eigenvalues of a Gram matrix which is constructed by evaluating a positive definite kernel function for each pair of data points.*

Proof. According to [31], the Rényi's α-order entropy of a random variable $\mathbf{X}$ can be defined as follows

$$H_\alpha(\mathbf{X}) = \frac{1}{1-\alpha} \log_2 \int_{\mathcal{X}} p^\alpha(x) dx \qquad (3)$$

where $\alpha \in (0,1) \cup (1,\infty)$, $p(x)$ is the probability density function of the random variable $\mathbf{X}$.

To better understanding, we take $\alpha = 2$ as an example to illustrate how the the eigenvalues of a Gram matrix fit the Rényi's 2-order entropy. For $\alpha = 2$, we can leverage the Parzen density estimator [17] with a Gaussian kernel $g_\sigma(x, y) = \exp\left(-\frac{1}{2\sigma^2}||x-y||^2\right)$ to calculate the probability density function $p(x)$, i.e., $\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} g_\sigma(x, x_i)$, which can be plugged into the Eq. 3, yields

$$
\begin{aligned}
\hat{H}_2(\mathbf{X}) &= -\log_2 \int_{\mathcal{X}} \hat{p}^2(x) dx \\
&= -\log_2 \int_{\mathcal{X}} \left(\frac{1}{n} \sum_{i=1}^{n} g_\sigma(x, x_i)\right)^2 dx \\
&= -\log_2 \left(\frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \int_{\mathcal{X}} g_\sigma(x, x_j) g_\sigma(x, x_i) dx\right) \\
&= -\log_2 \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g_\sigma(x_i, x_j)
\end{aligned}
\qquad (4)
$$

According to Eq. 4, the Rényi's 2-order entropy [11] of a random variable $\mathbf{X}$ can be characterized in terms of the eigenvalues of a Gram matrix G, where $G_{ij} = g_\sigma(x_i, x_j)$. Then, the Eq. 4 can be rewritten as

$$\hat{H}_2(\mathbf{X}) = -\log_2\left(\frac{1}{n^2}\text{tr}(G^T G)\right) \qquad (5)$$

The derivation of the Rényi's α-order entropy from its 2-order version can be found in Supplementary A, so $H_\alpha(\mathbf{X})$ can be calculated directly by the eigenvalues of a Gram matrix G which obtained by evaluating a positive definite kernel function g_σ for each pair of data points, but without considering intermediate steps in density estimation. $\square$

To facilitate the calculation of mutual information, we provide the definition of matrix-based Rényi's α-order entropy and joint entropy function as follows.

Definition 3 (Matrix-based Rényi's α-order entropy function). *Given a data collection $\mathbf{X}^v \in \mathcal{R}^{N \times D^v}$ from the v-th view, where N is the number of data samples. The Gram matrix $G_{\mathbf{x}}$ is obtained by calculating a positive definite kernel function $g_{\mathbf{x}}$ for all data pairs, i.e., $G_{\mathbf{x}}(i,j) = g_{\mathbf{x}}(x_i^v, x_j^v)$. The matrix-based Rényi's α-order entropy of X^v can be defined as follows*

$$H_\alpha(A_{\mathbf{x}}^v) = \frac{1}{1-\alpha} \log_2(tr((A_{\mathbf{x}}^v)^\alpha)) = \frac{1}{1-\alpha} \log_2\left(\sum_{i=1}^{N} \lambda_i(A_{\mathbf{x}}^v)^\alpha\right) \quad (6)$$

where $\alpha \in (0,1) \cup (1,\infty)$, $A_{\mathbf{x}}^v = \frac{G_{\mathbf{x}}}{tr(G_{\mathbf{x}})}$ is normalized from the Gram matrix $G_{\mathbf{x}}$, $\lambda_i(A_{\mathbf{x}}^v)$ indicates the i-th eigenvalue of $A_{\mathbf{x}}^v$.

Definition 4 (Matrix-based Rényi's α-order joint-entropy function). *Given a data collection $\mathbf{X}^v \in \mathcal{R}^{N \times D^v}$ from the v-th view and its corresponding representation $\mathbf{Z}^v \in \mathcal{R}^{N \times D^v}$, the matrix-based Rényi's α-order joint-entropy can be defined as follows*

$$H_\alpha(A_{\mathbf{x}}^v, A_{\mathbf{z}}^v) = H_\alpha\left(\frac{A_{\mathbf{x}}^v \circ A_{\mathbf{z}}^v}{tr(A_{\mathbf{x}}^v \circ A_{\mathbf{z}}^v)}\right) \qquad (7)$$

where $A_{\mathbf{x}}^v \circ A_{\mathbf{z}}^v$ indicates the Hadamard product between $A_{\mathbf{x}}^v$ and $A_{\mathbf{z}}^v$.

Given Eq. 6 and Eq. 7, the MI between high-dimensional variables can be directly calculated as follows

$$I_\alpha(\mathbf{X}^v; \mathbf{Z}^v) = H_\alpha(A_{\mathbf{x}}^v) + H_\alpha(A_{\mathbf{z}}^v) - H_\alpha(A_{\mathbf{x}}^v, A_{\mathbf{z}}^v) \qquad (8)$$

In next subsection, we prove that the novel MI measurement without variational approximation has analytical gradients that allow us to parameterize the IB principle and optimize it as an objective.

3.3. Deterministic Compression

IB aims to compress data observations and preserve the relevant information for a given task. Recently, it has been applied to analyze and understand the learning dynamics

of DNNs [34] benefiting by the progress of MI neural estimators, such as variational approximation [3]. However, existing MI neural estimators need the explicit estimation of the underlying distributions of data (more details can be found in Supplementary A),which often results in the uncertainty in representation learning. In this study, we design a novel MI measurement without variational approximation (Eq. 8), which is capable of directly fitting the MI of high-dimensional spaces. However, its differentiate property is unclear, which impedes its practical deployment as a loss function to parameterize IB principle. Next, we prove the MI measurement without variational approximation in Eq. 8 has an analytical gradient.

Proposition 2. *Given a data collection $X^v \in \mathcal{R}^{N \times D^v}$ from the v-th view and its corresponding representation $\mathbf{Z}^v \in \mathcal{R}^{N \times D^v}$, the mutual information measurement in Eq. 8 has an analytical gradient.*

Proof. First, we present the gradient of $H_\alpha(A_\mathbf{x}^v)$, which can be calculated as follows

$$\frac{\partial H_\alpha(A_\mathbf{x}^v)}{\partial A_\mathbf{x}^v} = \frac{\alpha}{1-\alpha} \frac{(A_\mathbf{x}^v)^{\alpha-1}}{(1-\alpha)tr((A_\mathbf{x}^v)^\alpha)} \tag{9}$$

Similarly, the gradient of $H_\alpha(A_\mathbf{x}^v, A_\mathbf{z}^v)$ can be calculated as follows

$$\frac{\partial H_\alpha(A_\mathbf{x}^v, A_\mathbf{z}^v)}{\partial A_\mathbf{x}^v} = \frac{\alpha}{1-\alpha} \left[\frac{(A_\mathbf{x}^v \circ A_\mathbf{z}^v)^{\alpha-1} \circ A_\mathbf{z}^v}{tr(A_\mathbf{x}^v \circ A_\mathbf{z}^v)^\alpha} - \frac{I \circ A_\mathbf{z}^v}{tr(A_\mathbf{x}^v \circ A_\mathbf{z}^v)} \right] \tag{10}$$

Since $I_\alpha(\mathbf{X}^v; \mathbf{Z}^v) = H_\alpha(A_\mathbf{x}^v) + H_\alpha(A_\mathbf{z}^v) - H_\alpha(A_\mathbf{x}^v, A_\mathbf{z}^v)$, $I_\alpha(\mathbf{X}^v; \mathbf{Z}^v)$ has an analytical gradient. □

In practice, the gradient of $I_\alpha(\mathbf{X}^v; \mathbf{Z}^v)$ can be computed using automatic differentiation libraries like Tensorflow and PyTorch.

In summary, based on **Proposition** 1, DIB can fit the mutual information from the original data and feature representation directly. Based on **Proposition** 2, the MI measurement without variational approximation has analytical gradients that allow us to parameterize the IB principle and optimize it as an objective. Thus, we construct a view-specific encoder E^v with analytical gradients for each view to parameterize the IB principle by directly fitting the MI $I_\alpha(\mathbf{X}^v; \mathbf{Z}^v)$ between the original view data $\{\mathbf{X}^v\}_{v=1}^V$ and representations $\{\mathbf{Z}^v\}_{v=1}^V$, i.e., $\mathbf{Z}^v = E^v(\mathbf{X}^v)$, which derives a deterministic compression of input variables from from different views. Note that, the MI measurement in Eq. 8 do not need any neural estimators to explicit estimate the underlying distribution of data, which enables us to parameterize IB principle with a deterministic neural network. Thus, we can obtain the loss function of deterministic compression inspired by IB principle as follows

$$\min \mathcal{L}_{comp} = \sum_{v=1}^V I_\alpha(\mathbf{X}^v; \mathbf{Z}^v) \tag{11}$$

3.4. Triplet Consistency Discovery

In MVC scenarios, another key issue is to capture the consistency across multiple views. Generally, multiple views of a data sample are different in attributes or input forms but show consistency in high-level features and semantics, which is also the foundation for effective MVC. Based on this observation, we first transform the compact representation of each view into high-level feature and cluster spaces by MLP. Then, a triplet consistency discovery mechanisms designed to mine the consistency across views from high-level features, clusters and joint of features and clusters.

Firstly, we design a **feature consistency** to make high-level features $\{\mathbf{H}^v\}_{v=1}^V$ focus on learning the common features across multiple views, which can be characterized through the popular contrastive learning [12, 28]. Specifically, each high-level feature h_i^v can form $(VN-1)$ feature pairs $\{h_i^v, h_j^m\}_{j=1,...,N}^{m=1,...,V}$ with all features except itself, where $\{h_i^v, h_i^m\}_{m \neq v}$ denotes $(V-1)$ positive feature pairs and $\{h_i^v, h_j^m\}_{j=1,...,N}^{m=1,...,V} - \{h_i^v, h_i^m\}_{m \neq v}$ denotes $V(N-1)$ negative feature pairs. In contrastive learning, the goal is to maximize the similarities between positive pairs while minimizing those of negative pairs. Then, the similarity between two features can be measured by the cosine distance as follows

$$d(h_i^v, h_j^m) = \frac{< h_i^v, h_j^m >}{||h_i^v|| \, ||h_j^m||} \tag{12}$$

where $< \cdot, \cdot >$ represents the dot product operator. And then, the feature consistency objective $\mathcal{L}_H$ between high-level features $\{\mathbf{H}^v\}_{v=1}^V$ can be formulated as

$$\max \mathcal{L}_{fea} = \sum_{v=1}^V \sum_{m \neq v} I(\mathbf{H}^v; \mathbf{H}^m)$$

$$\approx \sum_{v=1}^V \sum_{m \neq v} \mathbb{E} \left[\log \frac{e^{d(h_i^v, h_i^m)}}{\sum_{d(h_i^v, h_j^k) \in Neg}^N e^{d(h_i^v, h_j^k)}} \right] \tag{13}$$

$$+ V(V-1) \log N$$

where Neg denotes negative feature pairs, and $d(h_i^v, h_j^k)$ is the similarity of negative feature pairs.

Secondly, we can achieve the **cluster consistency** by contrastive learning to ensure the identical cluster labels convey consistent high-level semantics across views. Similarly, the cluster consistency objective $\mathcal{L}_{clu}$ between semantic labels $\{\mathbf{S}^v\}_{v=1}^V$ also can be calculated by Eq. 13 (the detailed formulation for $\mathcal{L}_{clu}$ can be found in Supplementary A).

Finally, we design a **joint consistency** between high-level features $\{\mathbf{H}^v\}_{v=1}^V$ and cluster assignments $\{\mathbf{S}^v\}_{v=1}^V$ to further refine the consistency across views. Intuitively, for one data instance, the learned feature representations from different views should maximally preserve the consistency for its cluster labels. This is to say, the mutual information between high-level features and cluster assignments also

Algorithm 1 Differentiable Information Bottleneck

1: **Input:** Multi-view data $\{\mathbf{X}^v\}_{v=1}^V$, cluster number K, iteration number I_t.
2: **Random initialization:** Initialize encoder E^v;
3: **for** epoch $\in \{0, 1, 2, \cdots, I_t\}$ **do**
4:     Obtain the representations $\{\mathbf{Z}^v\}_{v=1}^V$ via $\mathbf{Z}^v = E^v(\mathbf{X}^v)$.
5:     Obtain the high-level features $\{\mathbf{H}^v\}_{v=1}^V$ and semantic labels $\{\mathbf{S}^v\}_{v=1}^V$ through the feature MLP and label MLP, respectively.
6:     Calculate the feature, cluster consistency loss by Eq. 13 and calculate the joint consistency by Eq. 14.
7:     Calculate the compress loss function by Eq. 11.
8:     Update the parameters of the whole model by back propagation.
9: **end for**
10: **Output:** The clustering results $\mathbf{C}$.

should be maximized for better clustering. Thus, we define the joint consistency between high-level features $\{\mathbf{H}^v\}_{v=1}^V$ and cluster assignments $\{\mathbf{S}^v\}_{v=1}^V$ as follows

$$\max \mathcal{L}_{joint} = \sum_{v=1}^V I(\mathbf{H}^v; \mathbf{S}^v) \tag{14}$$

DIB consists of deterministic compression and triplet consistency discovery. Similar to IB principle, we construct a trade-off between deterministic compression and triplet consistency discovery as follows

$$\mathcal{L}_{overall} = \underbrace{\max(\mathcal{L}_{fea} + \mathcal{L}_{clu} + \gamma \mathcal{L}_{joint})}_{Consistency} + \beta \underbrace{\min \mathcal{L}_{comp}}_{Compression} \tag{15}$$

where γ and β are the trade-off parameters that control the impact of joint consistency and deterministic compression on the final clustering performance. $\mathcal{L}_{fea}$ and $\mathcal{L}_{clu}$ can be calculated by Eq. 13 as in contrastive clustering [28], $\mathcal{L}_{joint}$ and $\mathcal{L}_{comp}$ can be calculated by the proposed MI measurement without variational approximation as in Eq. 8. The DIB is presented in Algorithm 1.

4. Experiments

4.1. Datasets

The proposed DIB is evaluated on six widely-used multi-view datasets. MNIST-USPS [29] contains 5,000 samples of handwritten digit images based on two of features distributed across 10 categories. Berkeley drosophila genome project (BDGP) [4] contains 2,500 samples of Drosophila embryos belonging to 5 categories, each represented by visual and textual views. Handwritten [16] is a popular handwritten character dataset containing 2000 samples drawn by 6 different handwriting styles in 10 categories. Event segmentation and prediction (ESP) [36] is designed for action recognition, which captures 11,032 samples from 4 different viewpoints with different sensors or cameras, such as

RGB cameras and depth sensors. Flickr [7] is a widely used set of images that contains 12,154 samples from three shooting perspectives from different users at different locations and times, organized into seven categories. For Caltech [10], we construct three datasets with different numbers of views from on Caltech to evaluate the proposed method, i.e., Caltech-3V, Caltech-4V and Caltech-5V.

4.2. Implementation

The encoders in the DIB are composed of a four-layer fully connected network. The feature MLP consists of two linear layers. The label MLP is constructed by a linear layer and a Softmax layer. The DIB is implemented through Pytorch's public toolbox. We use the Adam optimizer for optimization and set the learning rate to 3×10^{-4}. The parameters α and β in the loss function (Eq. 15) are fixed, i.e., $\alpha = 0.01$ and $\beta = 0.01$, for all used datasets. We implement the experiments on a Windows 11 platform and an NVIDIA 4060Ti GPU with 16G of RAM.

4.3. Baselines

We compare the DIB with the three types of state-of-the-art methods. 1) Traditional MVC: binary MVC (BMVC) [50], simple multi-kernel k-means (SMKKM) [24], one-pass late fusion MVC (OPLFMVC) [23] and fast MVC via ensembles (FastMICE) [13]). 2) Deep MVC: multi-feature multi-level clustering (MFLVC) [43], cross-view contrastive learning (CVCL) [6], auto-weighted orthogonal and nonnegative graph reconstruction (AONGR) [51], global and cross-view feature aggregation (GCFAgg) [45] and self-discriminative MVC (SDMVC) [44]). 3) IB-based deep MVC: deep mutual information maximin (DMIM) [27], deep multi-view information bottleneck (DMIB) [9], consistency-guided multi-modal clustering (ConGMC) [28] and deep correlated information bottleneck (DCIB) [47]).

4.4. Evaluation Metrics

We use three widely-used clustering metrics including ACC [25], NMI [32] and purity (PUR) [2] to quantify the clustering results. All reported results are the average values by running 10 times.

4.5. Performance Analysis

We evaluate the effectiveness of the DIB with traditional, deep and IB-based MVC baselines. The comparison results are shown in Table 1. From this table, we obtain the following observations: 1) The DIB outperforms the traditional MVC, which demonstrates its superior ability of representation learning of high-dimensional space compared with traditional shallow MVC baselines. 2) Compared with several latest SOTA deep MVC baselines, DIB also achieves better performance. For example, the DIB

Table 1. Clustering performance of all methods on the six datasets. Bold and underline indicate the best and second best results.

Datasets	MNIST-USPS			BDGP			Handwritten			ESP		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
BMVC (TPAMI'2019)	72.22	60.37	73.50	85.68	71.98	85.68	84.45	77.59	84.45	47.95	32.01	50.33
SMKKM (ICCV'2021)	73.35	64.58	74.20	63.72	51.22	64.66	88.80	80.88	88.80	48.76	31.23	49.37
OPLFMVC (ICML'2021)	68.38	60.57	68.38	66.44	42.96	66.44	77.65	74.86	80.15	51.80	32.08	51.80
FastMICE (TKDE'2023)	90.73	90.24	91.44	77.42	62.81	77.91	84.55	86.68	85.78	54.94	36.32	55.29
MFLVC (CVPR'2022)	99.58	98.72	99.58	98.60	95.87	98.60	82.48	82.15	82.48	56.02	36.52	56.02
CVCL (ICCV'2023)	99.58	98.81	99.58	98.88	96.28	98.88	78.10	80.77	81.70	47.05	31.80	48.64
AONGR (INS'2023)	99.36	98.23	99.36	92.04	82.47	92.04	80.30	80.34	80.40	50.68	36.77	51.27
GCFAgg (CVPR'2023)	96.28	93.04	96.28	96.52	91.74	96.52	51.75	54.02	55.70	57.61	**40.59**	57.61
SDMVC (TKDE'2023)	99.82	99.47	99.82	96.80	92.00	96.80	77.63	86.92	77.63	49.57	36.16	49.57
DMIM (AAAI'2021)	98.12	97.53	98.09	93.18	92.63	93.49	81.23	83.74	82.83	56.11	37.26	55.42
DMIB (TCYB'2022)	96.71	97.12	97.35	96.57	95.20	96.32	80.92	81.66	81.15	51.03	23.17	50.68
ConGMC (TMM'2023)	99.01	98.45	98.62	97.28	94.36	95.51	83.65	84.29	84.57	58.45	37.62	58.37
DCIB (TNNLS'2023)	56.67	72.38	56.84	61.01	45.46	61.80	68.60	79.48	68.60	54.40	36.18	54.40
DIB (ours)	**99.86**	**99.56**	**99.86**	**99.00**	**96.65**	**99.00**	**88.95**	**89.92**	**88.95**	**59.06**	37.77	**59.06**

Datasets	Flicker			Caltech-3V			Caltech-4V			Caltech-5V		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
BMVC (TPAMI'2019)	56.73	36.04	56.80	64.93	53.46	45.38	73.36	69.74	73.36	77.29	70.96	77.29
SMKKM (ICCV'2021)	59.31	38.77	59.42	51.45	38.92	29.75	68.83	55.58	69.96	73.54	64.60	73.54
OPLFMVC (ICML'2021)	51.32	31.08	51.49	57.21	42.56	36.93	74.29	53.73	74.29	79.14	65.00	79.14
FastMICE (TKDE'2023)	54.75	35.50	54.99	64.23	57.25	50.19	73.90	66.88	77.00	78.63	72.48	78.63
MFLVC (CVPR'2022)	53.98	36.97	54.60	60.77	56.52	61.71	61.75	64.18	62.00	71.37	66.57	71.37
CVCL (ICCV'2023)	57.75	38.43	57.75	66.14	58.29	66.29	72.32	63.04	74.64	81.01	70.42	81.01
AONGR (INS'2023)	54.34	38.52	54.64	53.86	52.09	57.57	59.93	59.29	64.50	65.71	61.14	67.93
GCFAgg (CVPR'2023)	31.18	19.58	37.85	59.43	55.72	59.71	48.86	48.40	53.29	50.93	55.05	54.64
SDMVC (TKDE'2023)	39.30	19.04	39.31	67.66	57.72	50.52	74.79	68.03	77.79	83.84	78.08	83.84
DMIM (AAAI'2021)	57.44	34.83	57.68	70.71	58.67	69.34	73.07	70.94	74.09	79.28	63.09	80.16
DMIB (TCYB'2022)	55.74	27.88	56.29	71.21	59.23	70.52	72.78	66.82	71.95	82.28	68.02	81.97
ConGMC (TMM'2023)	60.05	37.92	57.06	73.37	64.80	74.54	73.78	69.14	75.25	83.78	73.55	82.70
DCIB (TNNLS'2023)	58.78	38.88	58.78	58.40	51.50	58.40	69.70	60.90	69.72	75.15	68.74	75.63
DIB (ours)	**60.40**	**40.23**	**60.40**	**74.71**	**68.46**	**75.71**	**75.64**	**71.51**	76.64	**84.79**	**78.34**	**84.79**

obtains 3.04%, 12.01%, 8.38%, 1.46% and 9.49% improvements compared with MFLVC, CVCL, AONGR, GCFAgg and SDMVC on the ESP dataset in terms of ACC metric. This is mainly because that the deterministic compression in DIB can learn a discriminative and compact representation for each view. 3) Compared with IB-based deep MVC baselines, DIB achieves the best results on all evaluation metrics in the used datasets. This is mainly because that the DIB can directly measure the information about feature representations from the source data rather than building a neural estimator to approximate the lower bound of mutual information. Besides, we conduct a significance test [33] to verify that the performance of the DIB is statistically better than the representative baselines (see Supplementary B for more details).

4.6. Ablation Study

To verify the effectiveness of each components, we consider the following five scenarios: A) Retain $\mathcal{L}_{clu}$. In this case, we only use the cluster consistency. B) Retain $\mathcal{L}_{clu}$ and $\mathcal{L}_{fea}$. We use the cluster consistency and feature consis-
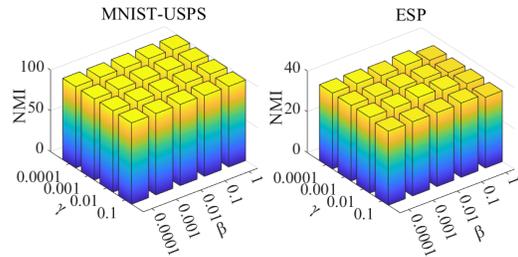

Figure 3. Parameter γ and β sensitivity experiment results.

tency simultaneously. C) Retain $\mathcal{L}_{clu}$, $\mathcal{L}_{fea}$, and $\mathcal{L}_{joint}$. In this scenario, we add joint consistency between cluster consistency and feature consistency so that the consistency learned from high-level features can further optimize the cluster structure. D) Retain $\mathcal{L}_{clu}$, $\mathcal{L}_{fea}$ and $\mathcal{L}_{comp}$. We add the deterministic compression component to scenario B). E) Retain $\mathcal{L}_{clu}$, $\mathcal{L}_{fea}$, $\mathcal{L}_{comp}$ and $\mathcal{L}_{joint}$. In this case, we perform MVC with the overall loss function of the DIB.

From Table 2, we can obtain the following observations. According to A) and B), we can find that learning

Table 2. Ablation experiments on loss components.

	Loss Components				MNIST-USPS			BDGP			ESP		
	$\mathcal{L}_{clu}$	$\mathcal{L}_{fea}$	$\mathcal{L}_{comp}$	$\mathcal{L}_{joint}$	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
A)	✓				85.72	90.24	85.72	82.58	87.13	82.58	48.01	31.96	49.43
B)	✓	✓			92.96	93.48	52.96	91.00	90.23	92.00	49.95	35.38	53.18
C)	✓	✓		✓	96.98	97.67	96.98	95.24	91.33	92.24	49.83	35.12	52.80
D)	✓	✓	✓		99.84	99.50	99.84	98.72	95.69	98.72	53.30	37.40	54.86
E)	✓	✓	✓	✓	**99.86**	**99.56**	**99.86**	**99.00**	**96.65**	**99.00**	**59.06**	**37.77**	**59.06**

Figure 4. Convergence curves on MNIST-USPS and ESP.

Figure 5. Mutual information with/without VA on the data pairs sampled from MNIST-USPS.
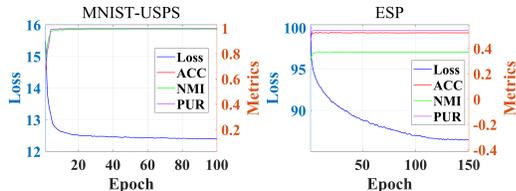
the consistency from high-level features can improve the clustering performance. This shows that it makes sense to map low-level features to high-level features to learn the common semantics. According to B) and D), we can observe that the deterministic compression can significantly improve the clustering performance. According to D) and E), we can find that establishing triplet consistency with the deterministic compression component can further enhance the clustering performance. This indicates that the learned deterministic and compact representation can facilitate the consistency discovery across multiple views. According to C) and D) and E), we can get that removing deterministic compression or discarding triplet consistency will make the final clustering performance degraded, which suggests that the two main parts of our model are highly integrated and refined. Besides, to further verify the effectiveness of the proposed MI measurement, we replace the differentiable MI measurement in DIB with the variational approximation (see Supplementary B for details).

4.7. Parameter Sensitivity Analysis

In this subsection, we evaluate the impact of the trade-off parameters γ and β on the clustering performance of the DIB on two representative datasets (MNIST-USPS and ESP). Specifically, we investigate the values of γ and β in the range of $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$, respectively. From Figure 3, we can observe that the DIB method can achieve stable clustering performance with different combinations of parameters in MNIST-USPS and ESP datasets. This suggests that our model is insensitive to the choice of γ and β. Based on the experimental results, we set $\gamma = 0.01$ and $\beta = 0.01$ for all used datasets in this study.
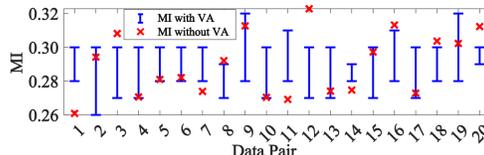
4.8. Convergence Analysis

Figure 4 reports the values of the loss function and the evaluation metrics of the DIB algorithm as the iterations increase. It can be observed that the loss function and evaluation metrics of the DIB approach to a fixed point with the epochs increase. This phenomenon shows that our DIB algorithm enjoys a good convergence property.

4.9. MI Measurement Evaluation

To verify the effectiveness of our MI without variational approximation (VA), we compare the MI with VA with our MI without VA by sampling 20 data pairs from MNIST-USPS dataset randomly. As shown in Figure 5, we observe that MI with VA fluctuates within a range and our MI without VA is a definite value.

5. Conclusions and Future Work

This paper investigates a novel differentiable information bottleneck method, which provides a deterministic and analytical MVC solution by essentially fitting the mutual information without the necessity of variational approximation. It is a valuable attempt to directly measure the information about feature representations from the data. In future, it is interesting to use the proposed mutual information measurement to conduct layer-by-layer training for a DNN.

6. Acknowledgments

References

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. 1, 4

[2] Ramiz M. Aliguliyev. Performance evaluation of density-based clustering methods. *Information Sciences*, 179(20): 3583–3602, 2009. 6

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *Proceeding of the International Conference on Machine Learning*, pages 530–539, 2018. 5, 1

[4] Xiao Cai, Hua Wang, Heng Huang, and Chris H. Q. Ding. Joint stage recognition and anatomical annotation of *drosophila* gene expression patterns. *Bioinformatics*, 28(12): 16–24, 2012. 6

[5] Jianpeng Chen, Yawen Ling, Jie Xu, Yazhou Ren, Shudong Huang, Xiaorong Pu, and Lifang He. Variational graph generator for multi-view graph clustering. *CoRR*, abs/2210.07011, 2022. 2

[6] Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. *arXiv preprint arXiv:2304.10769*, 2023. 1, 6

[7] Andrew M Cox, Paul D Clough, and Jennifer Marlow. Flickr: a first look at user behaviour in the context of photography as serious leisure. *Information Research*, 13(1):13–19, 2008. 6

[8] Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12350–12368, 2023. 1

[9] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020. 2, 6

[10] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 178–178, 2004. 6

[11] Luis Gonzalo Sánchez Giraldo, Murali Rao, and José C. Príncipe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2015. 4

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Coference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020. 5

[13] Dong Huang, Chang-Dong Wang, and Jian-Huang Lai. Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11388–11402, 2023. 6

[14] Shudong Huang, Yixi Liu, Ivor W. Tsang, Zenglin Xu, and Jiancheng Lv. Multi-view subspace clustering by joint measuring of consistency and diversity. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8270–8281, 2023. 1

[15] Zongmo Huang, Yazhou Ren, Xiaorong Pu, and Lifang He. Deep embedded multi-view clustering via jointly learning latent representations and graphs. *CoRR*, abs/2205.03803, 2022. 1, 3

[16] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. 6

[17] Robert Jenssen, Jose C Principe, Deniz Erdogmus, and Torbjørn Eltoft. The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006. 4, 1

[18] Aparajita Khan and Pradipta Maji. Multi-manifold optimization for multi-view subspace clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8): 3895–3907, 2022. 1

[19] Yunfan Li, Peng Hu, Jerry Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8547–8555, 2021. 3

[20] Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, and Zhaohua Yang. Deep adversarial multi-view clustering network. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2952–2958, 2019. 1, 3

[21] Renjie Lin, Shide Du, Shiping Wang, and Wenzhong Guo. Multi-channel augmented graph embedding convolutional network for multi-view clustering. *IEEE Transactions on Network Sciences and Engineering*, 10(4):2239–2249, 2023. 1, 3

[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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 12344–12353, 2021. 1

[23] Xinwang Liu, Li Liu, Qing Liao, Siwei Wang, Yi Zhang, Wenxuan Tu, Chang Tang, Jiyuan Liu, and En Zhu. One pass late fusion multi-view clustering. In *Proceedings of the International Conference on Machine Learning*, pages 6850–6859, 2021. 6

[24] Xinwang Liu, Sihang Zhou, Li Liu, Chang Tang, Siwei Wang, Jiyuan Liu, and Yi Zhang. Localized simple multiple kernel k-means. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9273–9281, 2021. 6

[25] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu, and Sen Wu. Understanding and enhancement of internal clustering validation measures. *IEEE Transactions on Cybernetics*, 43(3):982–994, 2013. 6

[26] Zhengzheng Lou, Yangdong Ye, and Xiaoqiang Yan. The multi-feature information bottleneck with application to unsupervised image categorization. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1508–1515, 2013. 1

[27] Yiqiao Mao, Xiaoqiang Yan, Qiang Guo, and Yangdong Ye. Deep mutual information maximin for cross-modal cluster-

ing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8893–8901, 2021. 1, 2, 6

[28] Yiqiao Mao, Xiaoqiang Yan, Jiaming Liu, and Yangdong Ye. Congmc: Consistency-guided multimodal clustering via mutual information maximin. *IEEE Transactions on Multimedia*, Early access:1–16, 2023. 1, 2, 5, 6

[29] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. COMIC: multi-view clustering without parameter selection. In *Proceedings of the International Conference on Machine Learning*, pages 5092–5101, 2019. 6

[30] José C. Príncipe, editor. *Information Theoretic Learning - Renyi's Entropy and Kernel Perspectives*. 2010. 3

[31] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–562, 1961. 4

[32] Alexander Strehl and Joydeep Ghosh. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002. 6

[33] Gan Sun, Yang Cong, Jiahua Dong, Yuyang Liu, Zhengming Ding, and Haibin Yu. What and how: Generalized lifelong spectral clustering via dual memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3895–3908, 2022. 7, 2

[34] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Transactions on Information Theory Workshop*, pages 1–5, 2015. 5

[35] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the Annual Allerton Conference on Communnication, Control and Computing*, pages 368–377, 1999. 1, 2

[36] Luis von Ahn and Laura Dabbish. ESP: labeling images with a computer game. In *Proceedings of the Knowledge Collection from Volunteer Contributors*, pages 91–98, 2005. 6

[37] Zhibin Wan, Changqing Zhang, Pengfei Zhu, and Qinghua Hu. Multi-view information-bottleneck representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10085–10092, 2021. 2

[38] Qianqian Wang, Zhiqiang Tao, Wei Xia, Quanxue Gao, Xiaochun Cao, and Licheng Jiao. Adversarial multiview clustering networks with adaptive fusion. *IEEE Transactions on Neural Networks Learning Systems*, 34(10):7635–7647, 2023. 1

[39] Shiye Wang, Changsheng Li, Yanming Li, Ye Yuan, and Guoren Wang. Self-supervised information bottleneck for deep multi-view subspace clustering. *CoRR*, abs/2204.12496, 2022. 2

[40] Shiye Wang, Changsheng Li, Yanming Li, Ye Yuan, and Guoren Wang. Self-supervised information bottleneck for deep multi-view subspace clustering. *IEEE Transactions on Image Processing*, 32:1555–1567, 2023. 1

[41] Chang Xu, Dacheng Tao, and Chao Xu. Large-margin multi-viewinformation bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1559–1572, 2014. 1

[42] 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. 3

[43] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16030–16039, 2022. 1, 3, 6

[44] Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, Philip S. Yu, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7470–7482, 2023. 6

[45] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 19863–19872, 2023. 1, 6

[46] Xiaoqiang Yan, Yangdong Ye, Xueying Qiu, and Hui Yu. Synergetic information bottleneck for joint multi-view and ensemble clustering. *Information Fusion*, 56:15–27, 2020. 1

[47] Xiaoqiang Yan, Yiqiao Mao, Yangdong Ye, and Hui Yu. Cross-modal clustering with deep correlated information bottleneck method. *IEEE Transactions on Neural Networks and Learning Systems*, Early access:1–15, 2023. 1, 2, 6

[48] Xiaoqiang Yan, Yingtao Gan, Yiqiao Mao, Yangdong Ye, and Hui Yu. Live and learn: Continual action clustering with incremental views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16264–16271, 2024. 1

[49] Xiaoqiang Yan, Yiqiao Mao, Mingyuan Li, Yangdong Ye, and Hui Yu. Multitask image clustering via deep information bottleneck. *IEEE Transactions on Cybernetics*, 54(3):1868–1881, 2024. 2

[50] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1774–1782, 2019. 6

[51] Mingyu Zhao, Weidong Yang, and Feiping Nie. Auto-weighted orthogonal and nonnegative graph reconstruction for multi-view clustering. *Information Sciences*, 632:324–339, 2023. 3, 6