Mutual Information-Driven Multi-View Clustering

Lei Zhang*
Sun Yat-Sen University
Guangzhou, China
zhanglei73@mail2.sysu.edu.cn

Lele Fu*
Sun Yat-Sen University
Guangzhou, China
fulle@mail2.sysu.edu.cn

Tong Wang
Sun Yat-Sen University
Guangzhou, China
wangt328@mail2.sysu.edu.cn

Chuan Chen†
Sun Yat-Sen University
Guangzhou, China
chenchuan@mail.sysu.edu.cn

Chuanfu Zhang
Sun Yat-Sen University
Guangzhou, China
zhangchf9@mail.sysu.edu.cn

ABSTRACT

In deep multi-view clustering, three intractable problems are posed ahead of researchers, namely, the complementarity exploration problem, the information preservation problem, and the cluster structure discovery problem. In this paper, we consider the deep multi-view clustering from the perspective of mutual information (MI), and attempt to address the three important concerns with a **M**utual **I**nformation-Driven **M**ulti-View **C**lustering (MIMC) method, which extracts the common and view-specific information hidden in multi-view data and constructs a clustering-oriented comprehensive representation. Specifically, three constraints based on MI are devised in response to three issues. Correspondingly, we minimize the MI between the common representation and view-specific representations to exploit the inter-view complementary information. Further, we maximize the MI between the refined data representations and original data representations to preserve the principal information. Moreover, to learn a clustering-friendly comprehensive representation, the MI between the comprehensive embedding space and cluster structure is maximized. Finally, we conduct extensive experiments on six benchmark datasets, and the experimental results indicate that the proposed MIMC outperforms other clustering methods.

CCS CONCEPTS

• **Information systems** → **Clustering**; • **Computing methodologies** → **Cluster analysis**.

KEYWORDS

Multi-view clustering, mutual information, contrastive learning.

ACM Reference Format:
Lei Zhang, Lele Fu, Tong Wang, Chuan Chen, and Chuanfu Zhang. 2023. Mutual Information-Driven Multi-View Clustering. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3583780.3614986

*Both authors contributed equally to this research.
†Corresponding author.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom.

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614986

1 INTRODUCTION

Multi-view clustering [10, 48, 50] (MVC) is an efficient data analysis technique derived from multi-view data, it deals with multiple views of data to extract the complementary information. A crucial step is to learn a consistent representation [4, 5, 8] that integrates the information across diverse views and reveals the discriminative quality of each sample. When it comes to multi-view data, its essence is the multi-source features of objects collected from varying domains. As a concrete example, relevant features can be extracted from an image in terms of color (e.g., Color Histogram [25]), texture (e.g., Local Binary Pattern [14]), etc. Semantic information from different perspectives might be heterogeneous, but it is a fundamental fact that various views conform to a uniform cluster distribution.

In the past few years, MVC field has been researched intensively, and a variety of traditional and deep algorithms with superior performance have been sequentially proposed, their concepts for exploring the view correlations are interesting and effective. In order to seek the consistency across views, some methods [3, 12, 28] mapped multi-source features into a unified subspace based on Canonical Correlation Analysis (CCA). How to balance contributions of different views is a critical problem, adaptive weight scheme is widely studied for this purpose. [20, 24, 33] integrated multiple graphs to match the consensus affinity matrix through an automatic weight assignment mechanism. Compared to mining the relationships among different views with a pairwise matrix pattern, some works [11, 18, 42] have been more concerned with excavating the high-order view correlations via a tensor-oriented manner. Recently, some multi-view representation learning works [19, 31, 34] based on information bottleneck theory have been proposed to filter superfluous information in multi-view data, thus preserving the critical information beneficial for downstream tasks.

In general, complementarity exploration and clustering structure discovery are two crucial concerns in both traditional and deep MVC. While the above methods have achieved considerable clustering results, they still need to be improved in these two aspects. For instance, the CCA-based and tensor-based methods [3, 11, 12, 18, 42] focus more on pursuing the consistency across views and ignore the role of view-specific information in enhancing the sample discrimination. Similarly, the weight-based approaches [20, 24, 33]

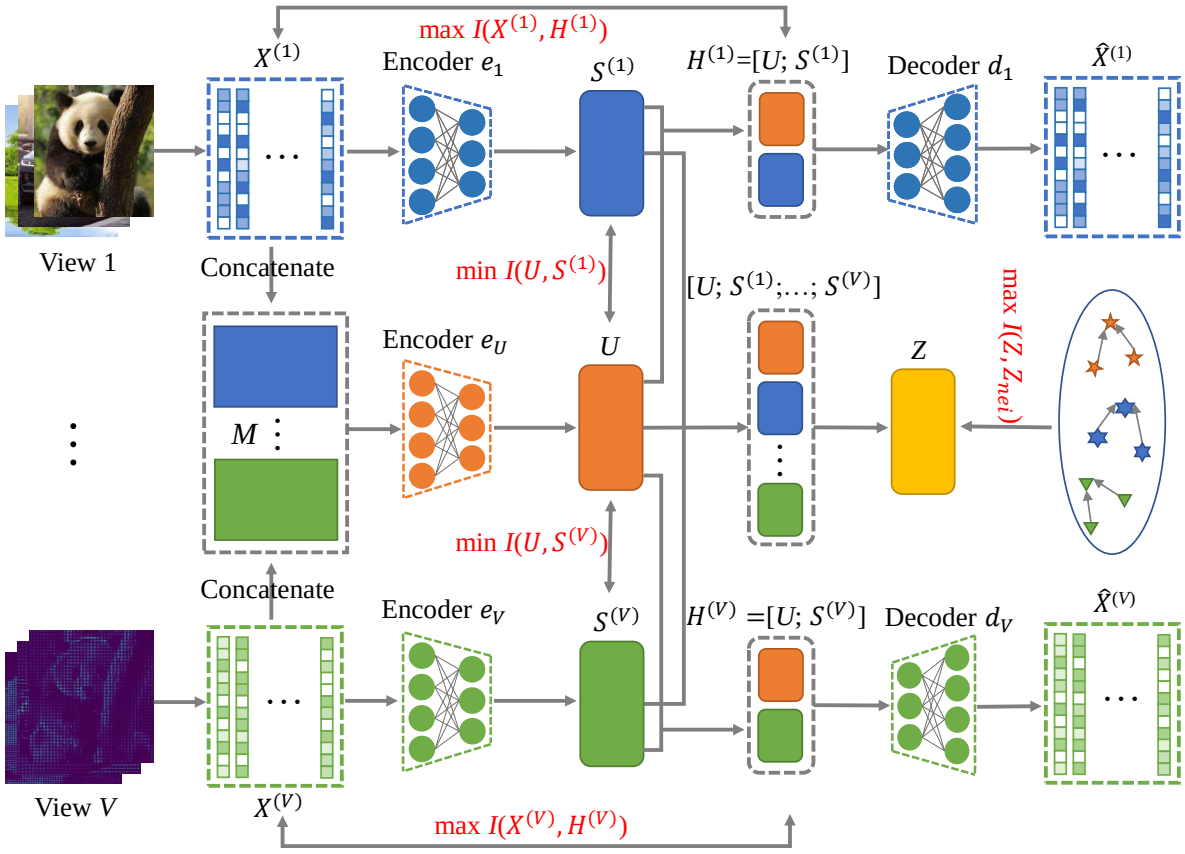


Figure 1: The general framework of the proposed MIMC. Specifically, the unified feature M is passed through the encoder e_U to yield the common representation U. Meanwhile, each feature representation $\mathrm{X}^{(v)}$ is input into its encoder e_v to acquire a compact representation $\mathrm{S}^{(v)}$. To obtain the multi-level data information, we minimize the MI between U and $\mathrm{S}^{(v)}$. Further, the v-th view's refined representation $\mathrm{H}^{(v)}$ is decoded via the decoder d_v to reconstruct the data $\hat{\mathrm{X}}^{(v)}$ with maximizing MI between $\mathrm{H}^{(v)}$ and $\mathrm{X}^{(v)}$. Considering the clustering task, we maximize the MI between instances and their k-nearest neighbors for enhancing intra-cluster aggregation. Overall, three constraints based on MI address the complementary exploration, information preservation, and cluster structure discovery, including $\min I(\mathrm{U}, \mathrm{S}^{(v)})$, $\max I(\mathrm{X}^{(v)}, \mathrm{H}^{(v)})$, **and** $\max I(\mathrm{Z}, \mathrm{Z}_{nei})$.

essentially focus on the view most similar to the consistent representation, which are limited for exploiting complementary information from all views. As for the current emerging multi-view representation learning methods [19, 31, 34] based on information bottleneck, they lack specific strategies for clustering tasks, and do not enhance the cluster structure of samples in the latent embedding space. Moreover, how to protect the principal information in the embedding space is an essential issue in deep MVC, limited works give an effective solution in terms of MI.

Motivated by these limitations, we provide a unified perspective from the MI for tackling the MVC problem, and propose a **M**utual **I**nformation-Driven **M**ulti-View **C**lustering (MIMC) method, which fully utilizes the common and view-specific information among multiple views and models a clustering-friendly comprehensive representation. Specifically, we use the common encoder and view-specific encoders to obtain a common representation and a set of view-specific representations, respectively. To acquire more information at different levels, the MI between the common representation and view-specific representations is constrained to be as small as possible. Further, we construct the refined representation composed of common and view-specific information for each view. With the guidance of preserving the critical information in the initial feature space, the refined representations are constrained to maximize the MI between the original representations via the view-specific decoders. Finally, the common representation and all view-specific representations are integrated together for modeling a comprehensive embedding space. Considering the clustering task, we expect that the sample points lied in the comprehensive embedding space are compliant with an explicit cluster structure. Then, a clustering-friendly representation is learned via maximizing the MI between the comprehensive embedding space and cluster structure. In general, the contributions of this paper are concluded as follows:

- We consider the deep MVC problem from the perspective of MI, and propose a unified framework to handle three important concerns, including the complementarity exploration, information preservation, and cluster structure discovery.
- We transform the hardly measurable MI estimation between representations into the form of optimizable loss function with the rigorous theoretical derivation.
- We conduct experiments on six multi-view datasets to validate the effectiveness of the proposed MIMC. Substantial experimental results demonstrate its superiority over the compared approaches.

We structure the remainder of this paper as follows. In Section 2, we review some dominant multi-view clustering methods and representation learning methods based on mutual information. In Section 3, the network architecture and the objective loss of the proposed method are introduced. Experimental settings and results are presented in Section 4. We summarize the paper in Section 5.

2 RELATED WORKS

2.1 Multi-view Clustering

Multifarious multi-view clustering methods are proposed in the past decades, we briefly introduce some representative ones herein. Nonnegative matrix factorization decomposes the target matrix into the form of multiplication of two nonnegative matrices, aiming at extracting the most significant feature elements. [15–17] explored a uniform nonnegative embedding matrix from multiple representations, which was a discrete variable and viewed as the label indicator. Subspace clustering is proficient at mining the underlying affinity relationship between instances and is broadly extended to multi-view scenarios. [20, 21, 49] were devoted to learning a well-structured subspace representation from multi-view data via information enhancement manners. Tensors [9, 45, 46] with multiple dimensions are naturally advantageous for capturing the intrinsic correlations of multi-source data. [13, 39, 44] leveraged the low-rank tensor approximation to recover the principal components of multi-view representation tensor. Deep learning is known for its sound ability to capture complex semantic information with the help of neural networks. Some auto-encoder based approaches [38, 41, 47] nonlinearly mapped multiple features into a compact subspace, wherein a clustering-friendly representation was further modeled via the Kullback-Leible divergence loss. While the above methods have achieved decent performance, the ability to mine multi-level information and enhance cluster structure in multi-view data needs further improvement. The proposed MIMC provides an effective framework for the multi-view clustering from a perspective of MI, excavating the multi-level information and strengthening the cohesiveness of samples within a same cluster via MI-based constraints.

2.2 Representation Learning with Mutual Information

Mutual information measures the amount of overlapping information between random variables. The greater the mutual information, the greater the correlations between variables, otherwise the smaller. Feature representations of data usually contain a multitude of superfluous information irrelevant to downstream tasks, then the mutual information theory is widely drawn upon for representation learning [6, 32, 37], with the objective of removing the redundant information. Mao et al. [22] explored the shared information across modalities via maximizing the MI between them. Schnapp et al. [26] selected important features with minimum MI with labels. In recent years, there are also some multi-view representation learning approaches incorporating MI theory emerging. Federici et al. [7] captured the shared information through maximizing the MI of representations under different views. Veyseh et al. [30] enhanced the semantic consensus between the sentence structures of two views via maximizing their MI. Wan et al. [31] learned a common representation and a series of view-specific representations based on the information bottleneck principle. Existing multi-view representation learning methods based on MI theory provide a novel schema for exploring view complementarity with desirable results, but they rarely incorporate feature learning strategies for the clustering tasks. On the contrary, the proposed MIMC models a clustering-friendly comprehensive representation via maximizing the MI between the embedding space and cluster structure.

3 THE PROPOSED METHOD

3.1 Network Architecture

Given a multi-view dataset $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$, where $\mathbf{X}^{(v)} \in \mathbb{R}^{d^{(v)} \times N}$ is the v-th feature matrix with N samples and $d^{(v)}$ dimension. The network architecture is composed of four modules: common encoder module, view-specific encoder module, view-specific decoder module, and neighbors enhancement module. Fig. 1 shows the overall framework of the proposed MIMC.

1) Common Encoder Module: To extract the common information, multiple feature matrices are concatenated into a unified representation $\mathbf{M}=[\mathbf{X}^{(1)}; \cdots; \mathbf{X}^{(V)}] \in \mathbb{R}^{\sum_{v=1}^{V} d^{(v)} \times N}$, which is further passed through the common encoder e_U to refine the common representation $\mathbf{U}$. Specifically, e_U is composed of three-layer fully connected layers. The encoding process is formulated as $\mathbf{U} = e_U(\mathbf{M}|\psi_{e_U})$, where ψ_{e_U} denotes the weights in the encoder e_U. The feature dimension is reduced from $\sum_{v=1}^{V} d^{(v)}$ to l after encoding.

2) View-Specific Encoder Module: Each view has its own view-specific information. To condense this information, we perform a nonlinear low-dimensional mapping using the view-specific encoder e_v over each $\mathbf{X}^{(v)}$, i.e., $\mathbf{S}^{(v)} = e_v(\mathbf{X}^{(v)}|\psi_{e_v})$, where e_v consists of three-layer fully connected layers, ψ_{e_v} is the weights of e_v. After encoding, the data dimension is transformed: $d^{(v)} \times N \rightarrow l \times N$. In order not to overlap the information embedded in the common representation and view-specific representations, we minimize the MI between the two to obtain multi-level information.

3) View-Specific Decoder Module: We define the refined view-specific representation as $\mathbf{H}^{(v)} = [\mathbf{U}; \mathbf{S}^{(v)}] \in \mathbb{R}^{2l \times N}$, the definition is based on the idea that view information contains the common information and view-specific information. The view-specific decoder d_v is designed with three-layer fully connected layers. $\mathbf{H}^{(v)}$ as input is passed through the decoder d_v to model the reconstructed data $\hat{\mathbf{X}}^{(v)}$, i.e., $\hat{\mathbf{X}}^{(v)} = d_v(\mathbf{H}^{(v)}|\psi_{d_v})$, where ψ_{d_v} denotes the parameters in d_v. In view of allowing $\mathbf{H}^{(v)}$ to retain as much critical information as possible in $\mathbf{X}^{(v)}$, the MI between the two is desired to be

maximized. Interestingly, it can be proved that enforcing $\hat{\mathbf{X}}^{(v)}$ to approximate $\mathbf{X}^{(v)}$ achieves this goal.

4) Neighbors Enhancement Module: The comprehensive representation $\mathbf{Z}$ containing common and all view-specific information is constructed by stacking $\mathbf{U}$ and $\{\mathbf{S}^{(v)}\}_{v=1}^{V}$. To discover the clustering structure, we attempt to enhance the MI between the embedding space and cluster structure. Specifically, the MI between each sample point $\mathbf{z}_i$ and its k-nearest neighbors $\{\mathbf{z}_j\}_{j\in\mathbb{N}_k}$ is maximized, thus achieving intra-cluster aggregation and inter-cluster separation in the overall. For the convenience of statement, we denote $\mathbf{Z}_{nei}$ as the k-nearest neighbor representation of $\mathbf{Z}$.

3.2 Objective Function

As mentioned above, three vital constraints based on MI in the proposed MIMC guarantee to learn a clustering-oriented comprehensive embedding space, including $\min I(\mathbf{U}, \mathbf{S}^{(v)})$, $\max I(\mathbf{X}^{(v)}, \mathbf{H}^{(v)})$, and $\max I(\mathbf{Z}, \mathbf{Z}_{nei})$, where $I(\cdot,\cdot)$ denotes the MI between two variables. Thus, our optimization goal is formulated as

$$\mathcal{L} = \sum_{v=1}^{V}\Big(\min I(\mathbf{U}, \mathbf{S}^{(v)}) + \max I(\mathbf{X}^{(v)}, \mathbf{H}^{(v)})\Big) + \max I(\mathbf{Z}, \mathbf{Z}_{nei}). \tag{1}$$

However, measuring the MI between variables are intractable. Fortunately, we can derive the MI constraints into three optimizable objective losses as follows.

1) $\mathbf{\min I(U, S^{(v)})}$: The common and view-specific representations are explored through common and view-specific encoders, respectively. We are concerned about a critical fact that the information included in the two representations may overlap excessively, and then the view complementarity is not fully mined. Therefore, it is expected to minimize the MI between them for facilitating the acquisition of multi-level data information. We formulate the MI between the common representation $\mathbf{U}$ and the view-specific representation $\mathbf{S}^{(v)}$ as

$$I(\mathbf{U}, \mathbf{S}^{(v)}) = \int\int p(\mathbf{u}, \mathbf{s}^{(v)}) log \frac{p(\mathbf{u}, \mathbf{s}^{(v)})}{p(\mathbf{u})p(\mathbf{s}^{(v)})} d\mathbf{u} d\mathbf{s}^{(v)}, \tag{2}$$

where $p(\cdot)$ indicates the probability density function. If the MI between $\mathbf{U}$ and $\mathbf{S}^{(v)}$ is expected to be minimal, the value of $p(\mathbf{u}, \mathbf{s}^{(v)})$ should be as similar to $p(\mathbf{u})p(\mathbf{s}^{(v)})$ as possible, which means that variable $\mathbf{u}$ and variable $\mathbf{s}^{(v)}$ are independent whenever possible. We adopt a frequently used independence measure function, i.e., the covariance function to assess the correlation between $\mathbf{u}$ and $\mathbf{s}^{(v)}$. In essence, we need to evaluate the correlation between the feature element $\mathbf{U}^i$ (the i-th row of $\mathbf{U}$) and the feature element $\mathbf{S}^{(v)j}$ (the j-th row of $\mathbf{S}^{(v)}$):

$$\begin{aligned} Cov(\mathbf{U}^i, \mathbf{S}^{(v)j}) &= \frac{1}{N}\sum_{n=1}^{N}(U_n^i - \mu_{\mathbf{U}}^i)(S_n^{(v)j} - \mu_{\mathbf{S}^{(v)}}^{(v)j}) \\ &= \frac{1}{N}(\mathbf{U}\mathbf{S}^{(v)^T})_{ij} - \mu_{\mathbf{U}}^i \mu_{\mathbf{S}^{(v)}}^{(v)j}, \end{aligned} \tag{3}$$

where U_n^i and $S_n^{(v)j}$ denote the n-th entry of $\mathbf{U}^i$ and $\mathbf{S}^{(v)j}$, respectively. $\mu_{\mathbf{U}}^i$ and $\mu_{\mathbf{S}^{(v)}}^{(v)j}$ denote the mean value of all entries in $\mathbf{U}^i$ and $\mathbf{S}^{(v)j}$, respectively. For the simplicity of optimization objective, we expect that $\mu_{\mathbf{U}}^i$ is equivalent to zero, which is readily achieved via $U_n^i = U_n^i - \mu_{\mathbf{U}}^i$. Thus, we only need to focus on the term $(\mathbf{U}\mathbf{S}^{(v)^T})_{ij}$. Our goal is to minimize the covariance of any $\mathbf{u}$ and $\mathbf{s}^{(v)}$, i.e., $\mathbf{U}\mathbf{S}^{(v)^T} \to \mathbf{0}$, where $\mathbf{0}$ is a matrix with all elements of zero. The problem $\mathbf{U}\mathbf{S}^{(v)^T} \to \mathbf{0}$ can be measured by $\min_{\mathbf{U},\mathbf{S}^{(v)}} ||\mathbf{U}\mathbf{S}^{(v)^T}||_0$. Since the minimization of l_0-norm is an NP hard problem, we relax l_0-norm to l_1-norm, and obtain a relaxed orthogonal loss form:

$$\mathcal{L}_{Ort} = \min_{\mathbf{U},\mathbf{S}^{(v)}} ||\mathbf{U}\mathbf{S}^{(v)^T}||_1. \tag{4}$$

2) $\mathbf{\max I(X^{(v)}, H^{(v)})}$: The refined representation $\mathbf{H}^{(v)}$ of the v-th view is obtained via $\mathbf{H}^{(v)} = [\mathbf{U}; \mathbf{S}^{(v)}]$. This construction manner follows the hypothesis that an individual view's representation consists of a common representation and a view-specific representation. Furthermore, we argue that a good refined representation retains the principal information in the initial feature space and is a compact formulation of the initial representation. Accordingly, we expect to maximize the MI between $\mathbf{H}^{(v)}$ and $\mathbf{X}^{(v)}$, which is written as

$$I(\mathbf{X}^{(v)}, \mathbf{H}^{(v)}) = \iint p(\mathbf{x}^{(v)}, \mathbf{h}^{(v)}) log(\frac{p(\mathbf{x}^{(v)}|\mathbf{h}^{(v)})}{p(\mathbf{x}^{(v)})}) d\mathbf{x}^{(v)} d\mathbf{h}^{(v)}. \tag{5}$$

For simplicity of formulas, we omit the superscript (v) in the following derivation. Let $q(\mathbf{x}|\mathbf{h})$ be the variational estimation of $p(\mathbf{x}|\mathbf{h})$, according to the definition of Kullback-Leibler (KL) divergence, we have

$$\begin{aligned} D_{KL}[p(\mathbf{x}|\mathbf{h}), q(\mathbf{x}|\mathbf{h})] &= \int p(\mathbf{x}|\mathbf{h}) log(\frac{p(\mathbf{x}|\mathbf{h})}{q(\mathbf{x}|\mathbf{h})}) d\mathbf{x} \geq 0 \\ \Rightarrow \int p(\mathbf{x}|\mathbf{h}) log(p(\mathbf{x}|\mathbf{h})) d\mathbf{x} &\geq \int p(\mathbf{x}|\mathbf{h}) log(q(\mathbf{x}|\mathbf{h})) d\mathbf{x}, \end{aligned} \tag{6}$$

then, we further have

$$\begin{aligned} &\int p(\mathbf{h}) d\mathbf{h} \int p(\mathbf{x}|\mathbf{h}) log(p(\mathbf{x}|\mathbf{h})) d\mathbf{x} \\ \geq &\int p(\mathbf{h}) d\mathbf{h} \int p(\mathbf{x}|\mathbf{h}) log(q(\mathbf{x}|\mathbf{h})) d\mathbf{x} \\ \Rightarrow &\int\int p(\mathbf{x},\mathbf{h}) log(\frac{p(\mathbf{x}|\mathbf{h})}{p(\mathbf{x})}) d\mathbf{x} d\mathbf{h} \\ \geq &\int\int p(\mathbf{x}, \mathbf{h}) log(\frac{q(\mathbf{x}|\mathbf{h})}{p(\mathbf{x})}) d\mathbf{x} d\mathbf{h}. \end{aligned} \tag{7}$$

From Eq. (7), the inequality related to $I(\mathbf{X}, \mathbf{H})$ can be obtained

$$\begin{aligned} I(\mathbf{X}, \mathbf{H}) \geq &\iint p(\mathbf{x}, \mathbf{h}) log(\frac{q(\mathbf{x}|\mathbf{h})}{p(\mathbf{x})}) d\mathbf{x} d\mathbf{h} \\ \geq &\iint p(\mathbf{x}, \mathbf{h}) log(q(\mathbf{x}|\mathbf{h})) d\mathbf{x} d\mathbf{h} - \iint p(\mathbf{x}, \mathbf{h}) log(p(\mathbf{x})) d\mathbf{x} d\mathbf{h}. \end{aligned} \tag{8}$$

Since $-\iint p(\mathbf{x}, \mathbf{h}) log(p(\mathbf{x})) d\mathbf{x} d\mathbf{h} \geq 0$, it further gives

$$I(\mathbf{X}, \mathbf{H}) \geq \int p(\mathbf{x}) d\mathbf{x} \int p(\mathbf{h}|\mathbf{x}) log(q(\mathbf{x}|\mathbf{h})) d\mathbf{h}. \tag{9}$$

According to Monte Carlo sampling method [27], we have

$$\int p(\mathbf{x}) d\mathbf{x} \int p(\mathbf{h}|\mathbf{x}) log(q(\mathbf{x}|\mathbf{h})) d\mathbf{h} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{p(\mathbf{h}|\mathbf{x}_i)} log(q(\mathbf{x}_i|\mathbf{h})). \tag{10}$$

Algorithm 1 The main steps of the proposed MIMC

Input: Multi-view data $\left\{\mathbf{X}^{(v)}\right\}_{v=1}^{V}$, parameters α, λ.
Output: Comprehensive representation $\mathbf{Z}$.

```
Select the Adam as the optimizer and initialize the learning rate to 0.001, the training epochs to 500.
for epoch = 1 to training epochs do
    Update e_U, {e_v}_{v=1}^V, and {d_v}_{v=1}^V via minimizing L_Rec via Eq. (14);
end for
for epoch = 1 to training epochs do
    Calculate the orthogonal loss L_Ort via Eq. (4);
    Calculate the reconstruction loss L_Rec via Eq. (14);
    Calculate the contrastive loss L_Con via Eq. (21);
    Calculate the complete objective loss L by Eq. (23) and update the network parameters via back propagation;
end for
Use Z as the input of K-means to yield the data labels.
```

Assuming that $q(\cdot)$ denotes the Gaussian density distribution, then $q(\mathbf{x}_i|\mathbf{h})$ has the following form:

$$q(\mathbf{x}_i|\mathbf{h}) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{||\mathbf{x}_i - R(\mathbf{h})||_2^2}{\sigma^2}). \tag{11}$$

Bringing Eq. (11) into Eq. (10), the following inequality holds:

$$I(\mathbf{X},\mathbf{H}) \geq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{p(\mathbf{h}|\mathbf{x}_i)}(\log(\sqrt{2\pi}\sigma) - \frac{||\mathbf{x}_i - R(\mathbf{h})||_2^2}{2\sigma^2}). \tag{12}$$

Hence, the problem of maximizing $I(\mathbf{X},\mathbf{H})$ is transformed into minimizing $\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{p(\mathbf{h}|\mathbf{x}_i)}||\mathbf{x}_i - R(\mathbf{h})||_2^2$. We use the Monte Carlo sampling method [27] to further simplify, then the following equation can be derived

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{p(\mathbf{h}|\mathbf{x}_i)}||\mathbf{x}_i - R(\mathbf{h})||_2^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M}||\mathbf{x}_i - R(\mathbb{H}_{i,m})||_2^2, \tag{13}$$

where $\mathbb{H}_{i,m}$ denotes the m-th distribution of the i-th refined sample. Specifically, when the data is encoded into the embedding space by a fixed encoder, M is equivalent to 1. Then we have $\mathbb{H}_{i,m} = \mathbf{h}_i$. According to above deduction, we can know that maximizing $I(\mathbf{X}^{(v)},\mathbf{H}^{(v)})$ is translated into the following minimization problem:

$$\mathcal{L}_{Rec} = \min_{\mathbf{h}_i^{(v)}}\sum_{i=1}^{N}||\mathbf{x}_i^{(v)} - R(\mathbf{h}_i^{(v)})||_2^2, \tag{14}$$

where $R(\cdot)$ can be viewed as the decoder d_v. Then, $R(\mathbf{h}_i^{(v)})$ is equivalent to the reconstructed data $\hat{\mathbf{x}}_i$. Hence, from the perspective of MI, the procedure of data reconstruction is essentially to maximize the MI between the refined representation $\mathbf{H}^{(v)}$ and the original data $\mathbf{X}^{(v)}$.

3) $\max I(\mathbf{Z},\mathbf{Z}_{nei})$*:* We form a comprehensive representation via $\mathbf{Z} = [\mathbf{U};\mathbf{S}^{(1)},\cdots,;\mathbf{S}^{(V)}]$. In unsupervised clustering scenario, there are no available data labels to guide the learning of feature representations, and it cannot be guaranteed that the learned representations are strongly correlated with their labels. Nevertheless, we instead consider to enhance the MI between the embedding space and cluster structure for modeling a clustering-oriented representation. Specifically, the MI between sample points and their k-nearest neighbors is expected to be maximized, thus achieving the intra-cluster aggregation and inter-cluster separation in a holistic context. The k-nearest neighbors enhancement is mathematically written as

$$\begin{aligned} I(\mathbf{Z},\mathbf{Z}_{nei}) &= \int\int p(\mathbf{z}_i,\mathbf{z}_j)log\frac{p(\mathbf{z}_j|\mathbf{z}_i)}{p(\mathbf{z}_j)}d\mathbf{z}_i d\mathbf{z}_j \\ &= \int p(\mathbf{z}_i)d\mathbf{z}_i \int p(\mathbf{z}_j|\mathbf{z}_i)log\frac{p(\mathbf{z}_j|\mathbf{z}_i)}{p(\mathbf{z}_j)}d\mathbf{z}_j, \end{aligned} \tag{15}$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ denote a variable in the comprehensive representation space and the neighborhood space of sample points, respectively. Using the Monte Carlo sampling, the following estimation can be made

$$I(\mathbf{Z},\mathbf{Z}_{nei}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_{nei}}\sum_{j=1}^{N_{nei}}log\frac{p(\mathbf{z}_j|\mathbf{z}_i)}{p(\mathbf{z}_j)}, \tag{16}$$

where N_{nei} is the number of neighbors. Thus, how to estimate $p(\mathbf{z}_j|\mathbf{z}_i)/p(\mathbf{z}_j)$ becomes the focus. Inspired by InfoNCE [29], we use a function $f(\mathbf{z}_j,\mathbf{z}_i)$ to model $p(\mathbf{z}_j|\mathbf{z}_i)/p(\mathbf{z}_j)$, and employ the cross entropy loss function to solve its optimal solution:

$$l_{CE} = -\log\frac{f(\mathbf{z}_j,\mathbf{z}_i)}{\sum_{k=1}^{N}f(\mathbf{z}_k,\mathbf{z}_i)}. \tag{17}$$

Theorem 1. *If the optimization objective Eq.* (17) *is minimized, then the mutual information between* $\mathbf{Z}$ *and* $\mathbf{Z}_{nei}$ *can be maximized.*

Proof. When estimating the overall cross entropy loss w.r.t. Eq. (17), we have

$$\begin{aligned} \mathcal{L}_{CE} &= -\frac{1}{N}\sum_{i}^{N}\sum_{j=1}^{N_{nei}}log\frac{f(\mathbf{z}_j,\mathbf{z}_i)}{\sum_{k=1}^{N}f(\mathbf{z}_k,\mathbf{z}_i)} \\ &= -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{N_{nei}}\left(log\frac{p(\mathbf{z}_j|\mathbf{z}_i)}{p(\mathbf{z}_j)} - log\sum_{k=1}^{N}\frac{p(\mathbf{z}_k|\mathbf{z}_i)}{p(\mathbf{z}_k)}\right), \end{aligned} \tag{18}$$

it can be derived that $log\sum_{k=1}^{N}\frac{p(\mathbf{z}_k|\mathbf{z}_i)}{p(\mathbf{z}_k)} = logN$. Hence, Eq. (18) is rewritten as

$$\begin{aligned} \mathcal{L}_{CE} &= -\frac{1}{N}\sum_{i}^{N}\sum_{j=1}^{N_{nei}}log\frac{p(\mathbf{z}_j|\mathbf{z}_i)}{p(\mathbf{z}_j)} + N_{nei}logN \\ &= -N_{nei}I(\mathbf{Z},\mathbf{Z}_{nei}) + N_{nei}logN. \end{aligned} \tag{19}$$

So far, we can observe that when $\mathcal{L}_{CE}$ reaches the minimum, $I(\mathbf{Z},\mathbf{Z}_{nei})$ takes the maximum. The proof is completed. □

According to Theorem 1, we only need to optimize Eq. (17) to the minimum, then achieve the goal of maximizing $I(\mathbf{Z},\mathbf{Z}_{nei})$.

To simplify calculation, we use sampling manner to estimate $\sum_{k=1}^{N}f(\mathbf{z}_k,\mathbf{z}_i)$, which is due to

$$\sum_{k=1}^{N_{sam}}f(\mathbf{z}_k,\mathbf{z}_i) = N_{sam}\mathbb{E}_k(f(\mathbf{z}_k,\mathbf{z}_i)) = \frac{N_{sam}}{N}\sum_{k=1}^{N}f(\mathbf{z}_k,\mathbf{z}_i), \tag{20}$$

where N_{sam} is the number of sampling data points. Thus, we can reformulate Eq. (18) into the form of current popular contrastive

Table 1: Statistics of six benchmark datasets.

Dataset ID	Datasets	Instances	Views	Clusters	Feature Dimensions	Category
1	ALOI	1079	4	10	64 / 64 / 77 / 13	Object image
2	GRAZ02	1476	6	4	512 / 32 / 256 / 500 / 500 / 680	Object image
3	MSRC	210	6	7	1302 / 48 / 512 / 100 / 256 / 210	Object image
4	Scene15	4485	3	15	1800 / 1180 / 1240	Scene image
5	UCI	2000	3	10	240 / 76 / 6	Digit image
6	WikipediaArticles	693	2	10	128 / 10	Document

learning:

$$\mathcal{L}_{Con} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_{pos}} log\frac{f(\mathbf{z}_j, \mathbf{z}_i)}{\sum_{k=1}^{N_{neg}} f(\mathbf{z}_k, \mathbf{z}_i)}. \tag{21}$$

It is worth noting that we consider the neighbors of the i-th data point as its positive instances, i.e., $N_{nei} = N_{pos}$, while the N_{sam} data points sampled in Eq. (20) are considered as its N_{neg} negative instances. We stipulate N_{neg} negative instances consist of all samples except the positive instances. Furthermore, the function $f(\mathbf{z}_j, \mathbf{z}_i)$ is defined as

$$f\left(\mathbf{z}_j, \mathbf{z}_i\right) = exp\left(\frac{\mathbf{z}_j^\top \mathbf{z}_i}{\left\|\mathbf{z}_j\right\| \left\|\mathbf{z}_i\right\|}\right). \tag{22}$$

Up to this point, we transform the intractable optimization goal Eq. (1) into the optimizable objective function:

$$\mathcal{L} = \mathcal{L}_{Rec} + \alpha\mathcal{L}_{Ort} + \lambda\mathcal{L}_{Con}, \tag{23}$$

where α and λ are two nonnegative hyperparameters to balance the three regularization terms. Algorithm 1 summarizes the flow of the proposed MIMC.

3.3 Training Process

The network training process can be divided into two phases: pre-training phase and complete training phase. We adopt the *Adam* optimizer and fix the learning rate as 0.001.

1) Pre-training phase: In pre-training phase, we send the stitched features $\mathbf{M}$ into the common encoder e_U to obtain the common representation $\mathbf{U}$, and pass $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$ through the view-specific encoders $\{e_v\}_{v=1}^{V}$ to get the view-specific representations $\{\mathbf{S}^{(v)}\}_{v=1}^{V}$, respectively. Thus, all the refined view-specific's representations $\{\mathbf{H}^{(v)}\}_{v=1}^{V}$ are input into multiple decoders $\{d_v\}_{v=1}^{V}$ to reconstruct the data $\{\hat{\mathbf{X}}^{(v)}\}_{v=1}^{V}$, respectively. Thus, we only calculate the reconstruction loss $\mathcal{L}_{Rec}$ between $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$ and $\{\hat{\mathbf{X}}^{(v)}\}_{v=1}^{V}$.

2) Complete training phase: The parameters of encoders and decoders are initialized with that obtained in pre-training. In the complete training phase, the reconstruction loss $\mathcal{L}_{Rec}$, orthogonal loss $\mathcal{L}_{Ort}$, and contrastive loss $\mathcal{L}_{Con}$ are all computed. Thus, the network parameters are further tuned via the back propagation. Finally, the proposed model yields the clustering-oriented comprehensive representation $\mathbf{Z}$. Our code is available at https://github.com/imvc2023/MIMC.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. We collect six public multi-view datasets to conduct experiments. Concretely, **ALOI**[1] is composed of 1079 object images in 10 categories and 4 views, including color histograms (CH), color similarities (CS), HSV color histograms, and Haralick features. **GRAZ02**[2] contains 1476 object images with 6 visual features, which are SURF feature, SIFT feature, LBP feature, WT feature, pyramid HOG, GIST feature, respectively. These images are divided into 4 classes. **MSRC**[3] includes 210 images divided into 7 categories, and 6 features are extracted such as HOG feature, LBP feature, SIFT feature, CENTRIST feature, GIST feature, and CM feature. **Scene15**[4] consists of 4485 scene images in 15 classes, each image has 3 features: PHOW feature, CENTRIST feature, and PRI-CoLBP feature. **UCI** [1] contains 2000 handwritten numeric images, the digits range from 0 to 9, 3 features are extracted: PIX feature, FOU feature, and MOR feature. **WikipediaArticles**[5] contains 693 short documents in 10 classes with 2 views. Table 1 provides a summary of main statistics of above datasets.

Baselines. We select the K-means algorithm as the basic baseline method and additionally collect ten SOTA multi-view clustering methods, including **AMGL** [23], **CSMSC** [21], **GMC** [33], **CGL** [18], **EOMSC-CA** [20], **MvDSCN** [49], **DSRL** [36], **DMSC-UDL** [35], **MFLVC** [43], **DFP-GNN** [40]. We set the parameters of these methods according to the values suggested in their paper. As for the proposed MIMC, we vary α in $\{0.0001, 0.0005\}$ and λ in $\{0.0005, 0.001, 0.01\}$, and fix the number of positive instances as 20. The dimensions of encoders are set as $d^{(v)} - 500 - 200 - 64$ or $d^{(v)} - 200 - 100 - 64$, and the decoders' dimensions are fixed as $128 - 200 - 500 - d^{(v)}$ or $128 - 100 - 200 - d^{(v)}$. ACC, NMI, ARI, F-score are used to evaluate the clustering performance. Each experiment is run 10 times and the average values are recorded.

4.2 Experimental Results and Analysis

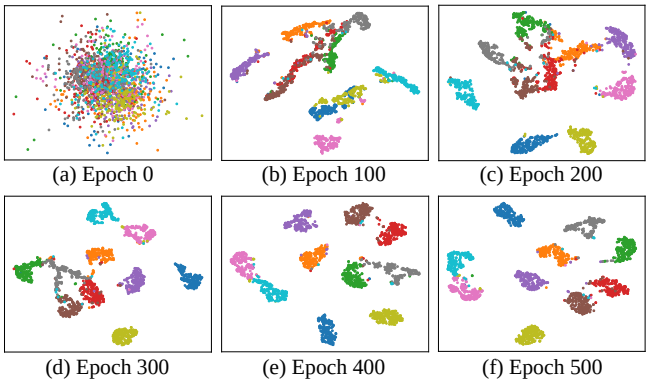
We illustrate the experimental results of the compared methods versus the proposed MIMC in Table 2. The best results are bolded, while the second best results are underlined. Traditional shallow models AMGL, CSMSC, GMC, CGL, EOMSC-CA perform feature transformations in a linear way, which do not efficiently cope with complex data distributions. Moreover, most of them only pursue

[1] https://elki-project.github.io/datasets/multi-view
[2] http://www.emt.tugraz.at/~pinz/data/GRAZ_02
[3] http://research.microsoft.com/en-us/projects/objectclassrecognition/
[4] http://www-cvr.ai.uiuc.edu/ponce_grp/data/
[5] http://lig-membres.imag.fr/grimal/data.html

Table 2: Clustering performance (%) of various methods on six datasets.

Methods	ALOI				GRAZ02				MSRC			
	ACC	NMI	ARI	F-score	ACC	NMI	ARI	F-score	ACC	NMI	ARI	F-score
K-means	47.49	47.34	32.98	41.04	35.91	3.2	3.56	33.83	45.00	38.08	24.28	37.39
AMGL	52.18	54.85	33.18	41.03	48.19	13.97	13.76	38.35	74.48	71.10	58.74	65.15
CSMSC	75.66	73.32	63.61	67.42	-	-	-	-	80.48	72.69	67.30	71.92
GMC	67.19	64.64	36.73	45.23	47.09	13.16	12.39	38.51	89.52	81.53	76.78	80.05
CGL	94.89	91.54	89.18	90.25	46.46	12.54	11.43	33.87	94.76	88.83	87.74	89.45
EOMSC-CA	65.80	76.36	47.24	54.49	42.48	12.19	10.66	33.33	69.52	71.34	57.70	64.10
MvDSCN	82.39	82.08	74.56	79.56	40.44	6.81	5.93	31.24	72.38	60.93	50.61	61.72
DSRL	78.47	78.71	61.71	64.41	51.98	17.38	17.08	38.92	85.76	78.79	69.86	74.31
DMSC-UDL	55.79	52.99	38.68	48.88	41.32	8.33	8.28	32.65	57.14	47.45	35.09	48.70
MFLVC	82.63	78.57	69.59	73.89	47.97	13.76	13.79	35.64	81.43	75.01	67.02	73.84
DFP-GNN	80.26	80.01	71.35	76.66	48.85	13.97	13.82	36.15	71.90	63.00	51.98	61.64
MIMC	**96.11**	**92.91**	**91.30**	**92.84**	**55.08**	**24.38**	**20.28**	**42.84**	**96.19**	**92.05**	**91.06**	**92.89**
Methods	Scene15				UCI				WikipediaArticles			
	ACC	NMI	ARI	F-score	ACC	NMI	ARI	F-score	ACC	NMI	ARI	F-score
K-means	30.49	28.79	15.00	21.93	38.76	46.64	31.35	38.86	54.69	51.48	39.02	45.80
AMGL	40.19	45.49	26.28	33.98	76.28	78.30	68.69	72.04	53.10	49.35	33.48	41.69
CSMSC	49.53	53.22	36.70	41.52	78.75	76.97	70.75	73.74	52.03	46.47	38.36	44.91
GMC	40.20	45.37	21.09	29.71	74.75	80.86	70.96	74.11	55.12	51.68	37.19	44.54
CGL	47.28	55.14	34.94	40.10	84.25	**90.52**	83.22	85.02	54.16	49.83	37.09	44.11
EOMSC-CA	45.13	52.40	30.98	37.11	54.80	67.09	46.28	53.57	56.13	52.91	42.30	48.47
MvDSCN	43.34	44.18	26.75	32.83	81.85	71.72	65.82	69.47	34.63	29.74	17.58	29.79
DSRL	10.74	0.58	0.11	10.73	78.28	79.99	70.57	73.77	59.06	50.10	39.19	47.24
DMSC-UDL	39.71	41.05	23.48	29.82	75.80	72.62	63.68	70.96	38.24	31.16	43.02	33.09
MFLVC	43.01	44.55	26.75	33.12	79.95	78.36	69.73	73.31	40.40	31.63	22.57	34.06
DFP-GNN	50.28	**55.44**	36.47	**43.81**	82.55	88.59	81.10	86.71	55.70	53.49	40.65	48.97
MIMC	**56.19**	54.67	**37.54**	43.74	**93.95**	90.49	**87.31**	**89.56**	**59.74**	**55.17**	**44.29**	**50.16**

the consensus across views. Nonetheless, MIMC leverages neural networks to map data nonlinearly into a compact space, and explores the common and view-specific information based on MI. In light of these, MIMC produces the clustering results surpassing that of traditional shallow methods. Compared to deep multi-view approaches MvDSCN, DSRL, DMSC-UDL, MFLVC, and DFP-GNN, MIMC is also in the leading position. DMSC-UDL employs the orthogonal regularization to obtain more data information, but its clustering effects are less effective than other methods, which may be attributed to difficulty of learning a consistent affinity matrix from multiple representations with enhanced diversity. To avoid this problem, MIMC concatenates the common and view-specific representations with their minimal MI to obtain a comprehensive representation. In the contrast strategy of MFLVC, only the same samples under different views are regarded as positive pairs with each other, the selection of positive instances is too restricted. In MIMC, k-nearest neighbors of a sample are recognized as it positive instances, which facilitates to produce intra-cluster aggregation effect and is more suitable for clustering task. Fig. 2 shows the gradual separation process of different clusters as the number of training epochs increases, it can be seen that different clusters are continuously pulled apart from each other, which means that sample features are increasingly discriminative.



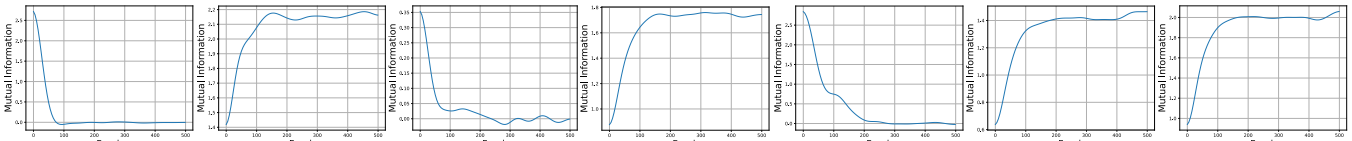
(a) Epoch 0 (b) Epoch 100 (c) Epoch 200

(d) Epoch 300 (e) Epoch 400 (f) Epoch 500

Figure 2: The scatter plot of MIMC's clustering results on UCI dataset as the training epoch increases via the t-SNE technology.

4.3 Mutual Information Evaluation

We adopt an efficient MI evaluator MINE [2] to measure the MI between different representations. Taking the UCI dataset as an example, Fig. 3 shows the value curves of MI between different representations with the increasing training epochs, including the



(a) $I(\mathrm{U}, \mathrm{S}^{(1)})$ (b) $I(\mathrm{X}^{(1)}, \mathrm{H}^{(1)})$ (c) $I(\mathrm{U}, \mathrm{S}^{(2)})$ (d) $I(\mathrm{X}^{(2)}, \mathrm{H}^{(2)})$ (e) $I(\mathrm{U}, \mathrm{S}^{(3)})$ (f) $I(\mathrm{X}^{(3)}, \mathrm{S}^{(3)})$ (g) $I(\mathrm{Z}, \mathrm{Z}_{\mathrm{nei}})$

Figure 3: The value curves of MI between representations as the training epoch increases on the UCI dataset. It can be seen that the change trends of MI conforms to the proposed optimization objectives.

Table 3: Ablation results (%) of the proposed MIMC on six datasets.

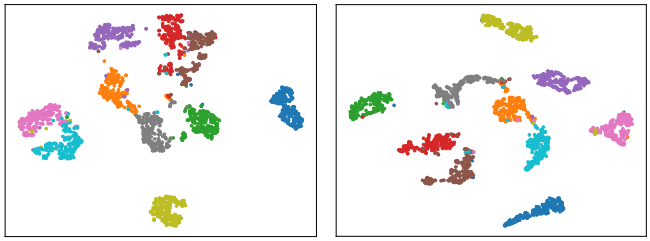
Loss	ALOI				GRAZ02				MSRC			
	ACC	NMI	ARI	F-score	ACC	NMI	ARI	F-score	ACC	NMI	ARI	F-score
$\mathcal{L}_{Rec}$	84.43 –	88.67 –	78.82 –	86.01 –	36.18 –	3.38 –	2.92 –	28.35 –	62.86 –	59.83 –	45.34 –	57.52 –
$\mathcal{L}_{Rec}$+$\mathcal{L}_{Ort}$	81.00 ↓	86.08 ↓	74.39 ↓	83.22 ↓	39.23 ↑	5.19 ↑	5.09 ↑	34.13 ↑	79.05 ↑	68.42 ↑	59.72 ↑	68.00 ↑
$\mathcal{L}_{Rec}$+$\mathcal{L}_{Con}$	91.66 ↑	88.40 ↑	83.74 ↑	86.17 ↑	48.98 ↑	17.25 ↑	14.54 ↑	37.38 ↑	74.29 ↑	73.01 ↑	59.67 ↑	71.79 ↑
$\mathcal{L}$	**96.11** ↑	**92.91** ↑	**91.30** ↑	**92.84** ↑	**55.08** ↑	**24.38** ↑	**20.28** ↑	**42.84** ↑	**96.19** ↑	**92.05** ↑	**91.06** ↑	**92.89** ↑
Loss	Scene15				UCI				WikipediaArticles			
	ACC	NMI	ARI	F-score	ACC	NMI	ARI	F-score	ACC	NMI	ARI	F-score
$\mathcal{L}_{Rec}$	21.76 –	22.83 –	7.90 –	19.61 –	81.15 –	79.74 –	71.57 –	75.04 –	40.12 –	33.53 –	18.02 –	32.88 –
$\mathcal{L}_{Rec}$+$\mathcal{L}_{Ort}$	50.93 ↑	53.63 ↑	34.76 ↑	42.55 ↑	82.80 ↑	79.95 ↑	73.60 ↑	76.77 ↑	42.71 ↑	44.65 ↑	26.38 ↑	38.86 ↑
$\mathcal{L}_{Rec}$+$\mathcal{L}_{Con}$	50.81 ↑	51.52 ↑	33.37 ↑	42.19 ↑	83.40 ↑	81.73 ↑	74.73 ↑	77.97 ↑	48.63 ↑	45.19 ↑	30.72 ↑	40.76 ↑
$\mathcal{L}$	**56.19** ↑	**54.67** ↑	**37.54** ↑	**43.74** ↑	**93.95** ↑	**90.49** ↑	**87.31** ↑	**89.56** ↑	**59.74** ↑	**55.17** ↑	**44.29** ↑	**50.16** ↑

MI between the common representation $\mathbf{U}$ and view-specific representation $\mathbf{S}^{(v)}$, the MI between the initial representation $\mathbf{X}^{(v)}$ and refined representation $\mathbf{H}^{(v)}$, and the MI between the comprehensive representation $\mathbf{Z}$ and neighbor representation $\mathbf{Z}_{nei}$. It can be seen that the $I(\mathbf{U}, \mathbf{S}^{(v)})$ drops rapidly while the $I(\mathbf{X}^{(v)}, \mathbf{H}^{(v)})$ and $I(\mathbf{Z}, \mathbf{Z}_{nei})$ rise rapidly, which demonstrates that the derived loss functions can achieve the corresponding optimization objectives of MI between representations.

4.4 Ablation Study

The proposed MIMC consists of three essential losses, including orthogonal loss $\mathcal{L}_{Ort}$, reconstruction loss $\mathcal{L}_{Rec}$, and contrastive loss $\mathcal{L}_{Con}$. Table 3 exhibits the clustering results when MIMC is equipped with different losses. We can observe that when MIMC has only $\mathcal{L}_{Rec}$, the clustering performance is almost always the worst, and it is improved by coupling $\mathcal{L}_{Rec}$ with $\mathcal{L}_{Ort}$ or $\mathcal{L}_{Con}$. Certainly, the best results are obtained with using the three losses in combination.

Fig. 4 shows the scatter plots with different losses on the UCI dataset. When MIMC contains only $\mathcal{L}_{Rec}$ (Fig. 4(a)), different clusters are close together without sound separability. After $\mathcal{L}_{Ort}$ is introduced (Fig. 4(b)), the division of instances is somewhat improved. The main reason is that multi-level data information is obtained and the data discrimination is enhanced. Nevertheless, the overall separability still needs to be raised. Interestingly, as MIMC is equipped with $\mathcal{L}_{Con}$ (Fig. 4(c)), although the separability of diverse clusters is boosted, the division of some samples is apparently not reasonable. When MIMC is armed with three losses (Fig. 4(d)), it can be observed that the overall isolation of clusters and the rationality of segmentation is the best.



(a) L_{Rec} (b) $L_{Rec} + L_{Ort}$ (c) $L_{Rec} + L_{Con}$ (d) L

Figure 4: Scatter plot with different losses of the proposed MIMC on UCI dataset via the t-SNE technology.

In addition, to verify the contribution of view-specific information to boost multi-view clustering performance, we perform

clustering with the common representation and common+view-specific representation, respectively. The performance comparison is illustrated in Fig. 5. It can be noticed that using only the common representation gives worse results than using the common+view-specific representation on all datasets, which suggests that view-specific information is beneficial to improve the clustering effects.

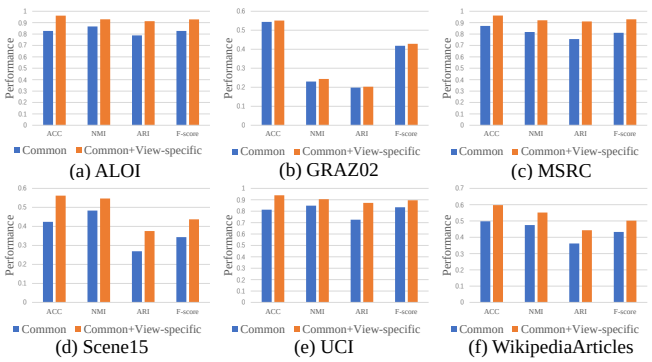


Figure 5: Clustering performance using the common representation and the common+view-specific representation, respectively.

4.5 Parameter Sensitivity Investigation

In the objective function Eq. (23), there are two trade-off parameters α and λ. To inquire into their effects on the clustering performance of the proposed model, we vary α and λ in {0.0001, 0.001, 0.01, 0.1, 1, 10}, and report the experimental results with diverse parameter combinations in Fig. 6. It can be seen that the clustering performance is inferior when α is set to a large value, which may be attributed to excessive penalty for the orthogonal loss $\mathcal{L}_{Ort}$, resulting in too many zero-elements in $\mathbf{U}$ or $\mathbf{S}^{(v)}$ and the missing of substantial information. Significantly, when λ is tuned to be relatively large, the performance turns a little better, which thanks to more importance given to the contrastive loss $\mathcal{L}_{Con}$, thus facilitating the cluster division.

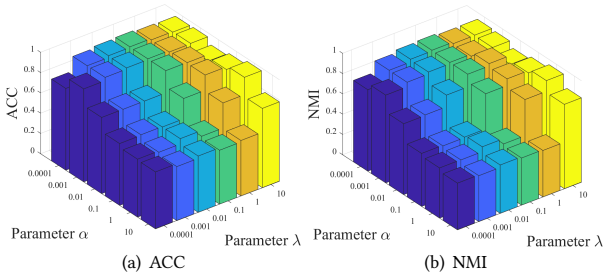


Figure 6: Parameter sensitivity analysis on ALOI dataset.

4.6 Convergence Analysis

We verify the convergence of the proposed MIMC in the Fig. 7. The horizontal axis represents the number of training epochs and the vertical axis represents the values of loss $\mathcal{L}$. As we can see, the loss value rapidly decreases at first and then gradually stabilizes with continuous training. Therefore, the convergence of the proposed MIMC is guaranteed.

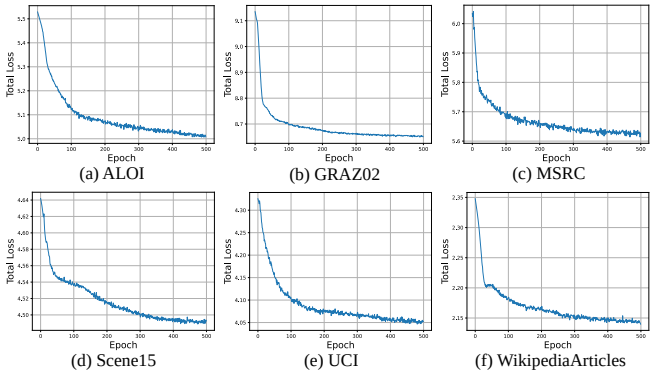


Figure 7: Convergence curves of the proposed MIMC on six datasets.

5 CONCLUSION

In this paper, we propose a novel multi-view clustering approach from the perspective of MI. Three constraints of MI are considered to model a clustering-friendly representation. Specifically, to obtain multi-level information, the MI of common representation and view-specific representation is forced to be minimized. Thus, we reconstruct the original data from the refined representation with guaranteeing the maximization of their MI. Furthermore, in order to make the comprehensive representation more suitable for clustering task, the MI of the comprehensive feature space and cluster structure is maximized. Finally, numerous experimental results on six challenging multi-view datasets confirm the effectiveness of the proposed MIMC.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (62176269, U1911202), the Guangzhou Science and Technology Program (2023A04J0314), and the Sun Yat-Sen University Young Faculty Development Program (23ptpy109).

REFERENCES

[1] Arthur Asuncion and David Newman. 2007. UCI machine learning repository.
[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *Proceedings of the International Conference on Machine Learning*. 531–540.
[3] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. 2009. Multi-view clustering via canonical correlation analysis. In *Proceedings of the International Conference on Machine Learning*. 129–136.
[4] Yongyong Chen, Xiaolin Xiao, and Yicong Zhou. 2020. Multi-view subspace clustering via simultaneously learning the representation tensor and affinity matrix. *Pattern Recognition* 106 (2020), 107441.
[5] Zhaoliang Chen, Lele Fu, Jie Yao, Wenzhong Guo, Claudia Plant, and Shiping Wang. 2023. Learnable graph convolutional network and feature fusion for multi-view learning. *Information Fusion* 95 (2023), 109–119.
[6] Jinfu Fan, Yang Yu, Linqing Huang, and Zhongjie Wang. 2023. GraphDPI: Partial label disambiguation by graph representation learning via mutual information maximization. *Pattern Recognition* 134 (2023), 109133.

[7] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2020. Learning Robust Representations via Multi-View Information Bottleneck. In *Proceedings of the International Conference on Learning Representations*.

[8] Lele Fu, Zhaoliang Chen, Yongyong Chen, and Shiping Wang. 2022. Unified Low-Rank Tensor Learning and Spectral Embedding for Multi-View Subspace Clustering. *IEEE Transactions on Multimedia* (2022). doi: 10.1109/TMM.2022.3185886.

[9] Lele Fu, Zhaoliang Chen, Sujia Huang, Sheng Huang, and Shiping Wang. 2021. Multi-View Learning Via Low-Rank Tensor Optimization. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. 1–6.

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

[11] Lele Fu, Jinghua Yang, Chuan Chen, and Chuanfu Zhang. 2022. Low-rank tensor approximation with local structure for multi-view intrinsic subspace clustering. *Information Sciences* 606 (2022), 877–891.

[12] Quanxue Gao, Huanhuan Lian, Qianqian Wang, and Gan Sun. 2020. Cross-Modal Subspace Clustering via Deep Canonical Correlation Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3938–3945.

[13] Jipeng Guo, Yanfeng Sun, Junbin Gao, Yongli Hu, and Baocai Yin. 2023. Logarithmic Schatten-pp Norm Minimization for Tensorial Multi-View Subspace Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2023), 3396–3410.

[14] Zhenhua Guo, Lei Zhang, and David Zhang. 2010. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing* 19, 6 (2010), 1657–1663.

[15] S. El Hajjar, Fadi Dornaika, and Fahed Abdallah. 2022. Multi-view spectral clustering via constrained nonnegative embedding. *Information Fusion* 78 (2022), 209–217.

[16] Zhanxuan Hu, Feiping Nie, Rong Wang, and Xuelong Li. 2020. Multi-view spectral clustering via integrating nonnegative embedding and spectral embedding. *Information Fusion* 55 (2020), 251–259.

[17] Chenglu Li, Hangjun Che, Man-Fai Leung, Cheng Liu, and Zheng Yan. 2023. Robust multi-view non-negative matrix factorization with adaptive graph and diversity constraints. *Information Sciences* 634 (2023), 587–607.

[18] Zhenglai Li, Chang Tang, Xinwang Liu, Xiao Zheng, Wei Zhang, and En Zhu. 2022. Consensus Graph Learning for Multi-View Clustering. *IEEE Transactions on Multimedia* 24 (2022), 2461–2472.

[19] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. 2023. Dual Contrastive Prediction for Incomplete Multi-View Representation Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2023), 4447–4461.

[20] Suyuan Liu, Siwei Wang, Pei Zhang, Xinwang Liu, Kai Xu, Changwang Zhang, and Feng Gao. 2022. Efficient One-pass Multi-view Subspace Clustering with Consensus Anchors. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[21] Shirui Luo, Changqing Zhang, Wei Zhang, and Xiaochun Cao. 2018. Consistent and Specific Multi-view Subspace Clustering. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*. 3730–3737.

[22] Yiqiao Mao, Xiaoqiang Yan, Qiang Guo, and Yangdong Ye. 2021. Deep Mutual Information Maximin for Cross-Modal Clustering. *Proceedings of the AAAI Conference on Artificial Intelligence* (2021), 8893–8901.

[23] Feiping Nie, Jing Li, and Xuelong Li. 2016. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*. 1881–1887.

[24] Feiping Nie, Jing Li, and Xuelong Li. 2017. Self-weighted Multiview Clustering with Multiple Graphs. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*. 2564–2570.

[25] Carol L Novak, Steven A Shafer, et al. 1992. Anatomy of a color histogram.. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 599–605.

[26] Shachar Schnapp and Sivan Sabato. 2021. Active Feature Selection for the Mutual Information Criterion. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9497–9504.

[27] Alexander Shapiro. 2003. Monte Carlo sampling methods. *Handbooks in operations research and management science* 10 (2003), 353–425.

[28] Jianqin Sun, Xianchao Xiu, Ziyan Luo, and Wanquan Liu. 2023. Learning High-Order Multi-View Representation by New Tensor Canonical Correlation Analysis. *IEEE Transactions on Circuits and Systems for Video Technology* (2023). doi:10.1109/TCSVT.2023.3263853.

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* (2019).

[30] Amir Pouran Ben Veyseh, Franck Dernoncourt, My Tra Thai, Dejing Dou, and Thien Huu Nguyen. 2020. Multi-View Consistency for Relation Extraction via Mutual Information and Structure Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9106–9113.

[31] Zhibin Wan, Changqing Zhang, Pengfei Zhu, and Qinghua Hu. 2021. Multi-View Information-Bottleneck Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 10085–10092.

[32] Feng Wang, Tao Kong, Rufeng Zhang, Huaping Liu, and Hang Li. 2023. Self-Supervised Learning by Estimating Twin Class Distribution. *IEEE Transactions on Image Processing* 32 (2023), 2228–2236.

[33] H. Wang, Y. Yang, and B. Liu. 2020. GMC: Graph-Based Multi-View Clustering. *IEEE Transactions on Knowledge and Data Engineering* 32, 6 (2020), 1116–1129.

[34] Qi Wang, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou. 2019. Deep Multi-view Information Bottleneck. In *Proceedings of the International Conference on Data Mining*. 37–45.

[35] Qianqian Wang, Jiafeng Cheng, Quanxue Gao, Guoshuai Zhao, and Licheng Jiao. 2021. Deep Multi-View Subspace Clustering With Unified and Discriminative Learning. *IEEE Transactions on Multimedia* 23 (2021), 3483–3493.

[36] Shiping Wang, Zhaoliang Chen, Shide Du, and Zhouchen Lin. 2022. Learning Deep Sparse Regularizers With Applications to Multi-View Clustering and Semi-Supervised Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2022), 5042–5055.

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

[38] Jie Wen, Zhihao Wu, Zheng Zhang, Lunke Fei, Bob Zhang, and Yong Xu. 2021. Structural Deep Incomplete Multi-view Clustering Network. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 3538–3542.

[39] Wei Xia, Quanxue Gao, Qianqian Wang, Xinbo Gao, Chris Ding, and Dacheng Tao. 2023. Tensorized Bipartite Graph Learning for Multi-View Clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 4 (2023), 5187–5202.

[40] Shunxin Xiao, Shide Du, Zhaoliang Chen, Yunhe Zhang, and Shiping Wang. 2023. Dual Fusion-Propagation Graph Neural Network for Multi-View Clustering. *IEEE Transactions on Multimedia* (2023), 1–13. doi=10.1109/TMM.2023.3248173.

[41] Yuan Xie, Bingqian Lin, Yanyun Qu, Cuihua Li, Wensheng Zhang, Lizhuang Ma, Yonggang Wen, and Dacheng Tao. 2021. Joint Deep Multi-View Learning for Image Clustering. *IEEE Transactions on Knowledge and Data Engineering* 33, 11 (2021), 3594–3606.

[42] Yuan Xie, Dacheng Tao, Wensheng Zhang, Yan Liu, Lei Zhang, and Yanyun Qu. 2018. On Unifying Multi-view Self-Representations for Clustering by Tensor Multi-rank Minimization. *International Journal of Computer Vision* 126, 11 (2018), 1157–1179.

[43] Jie Xu, Huayi Tang, Yazhou Ren, Xiaofeng Zhu, and Lifang He. 2022. Multi-level Feature Learning for Contrastive Multi-view Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 16051–16060.

[44] Jing-Hua Yang, Chuan Chen, Hong-Ning Dai, Meng Ding, Le-Le Fu, and Zibin Zheng. 2022. Hierarchical Representation for Multi-View Clustering: From Intra-Sample to Intra-View to Inter-View. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 2362–2371.

[45] Jing-Hua Yang, Chuan Chen, Hong-Ning Dai, Meng Ding, Zhe-Bin Wu, and Zibin Zheng. 2022. Robust corrupted data recovery and clustering via generalized transformed tensor low-rank representation. *IEEE Transactions on Neural Networks and Learning Systems* (2022). doi:.

[46] Jing-Hua Yang, Chuan Chen, Hong-Ning Dai, Le-Le Fu, and Zibin Zheng. 2022. A structure noise-aware tensor dictionary learning method for high-dimensional data clustering. *Information Sciences* 612 (2022), 87–106.

[47] Changqing Zhang, Yeqing Liu, and Huazhu Fu. 2019. AE2-Nets: Autoencoder in Autoencoder Networks. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*. 2577–2585.

[48] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. 2017. Multi-view learning overview: Recent progress and new challenges. *Information Fusion* 38 (2017), 43–54.

[49] Pengfei Zhu, Binyuan Hui, Changqing Zhang, Dawei Du, Longyin Wen, and Qinghua Hu. 2019. Multi-view Deep Subspace Clustering Networks. *ArXiv: 1908.01978* (2019).

[50] Linlin Zong, Faqiang Miao, Xianchao Zhang, and Bo Xu. 2020. Multimodal Clustering via Deep Commonness and Uniqueness Mining. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 2357–2360.