Multi-aspect Self-guided Deep Information Bottleneck for Multi-modal Clustering

Shizhe Hu, Jiahao Fan, Guoliang Zou*, Yangdong Ye *

School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China
ieshizhehu@gmail.com, iejhfan@gs.zzu.edu.cn, jimmyopop8@gmail.com, ieydye@zzu.edu.cn

Abstract

Deep multi-modal clustering can extract useful information among modals, thus benefiting the final clustering and many related fields. However, existing multi-modal clustering methods have two major limitations. First, they often ignore different levels of guiding information from both the feature representations and cluster assignments, which thus are difficult in learning discriminative representations. Second, most methods fail to effectively eliminate redundant information between multi-modal data, negatively affecting clustering results. In this paper, we propose a novel multi-aspect self-guided deep information bottleneck (MSDIB) method for multi-modal clustering, which can effectively employ different aspects of guiding information for learning cluster-friendly information among modals. MSDIB mainly contains two parts: information compression and information preservation. In information compression, we extract from the private information of each modality to obtain the compact representation and meanwhile conduct mutual compression between them. In information preservation, the aim is to preserve the shared information among modals and the self-supervised information from the clustering results in each iteration. In the above process, there are mainly three aspects of self-guiding information, the modality-private information, the modality-shared information and the self-supervised pseudo label information. By minimizing the mutual information based objective function with a variational optimization method, we can fully extract useful discriminative information while eliminating the irrelevant parts. Extensive experimental results demonstrate that our method outperforms state-of-the-art multi-modal clustering methods, showcasing its superior performance and broad application prospects.

Code — https://github.com/ShizheHu

Introduction

In real-world, objects usually exhibit rich modalities, where the information is conveyed by diversified media, including images, text descriptions or videos. Multi-modal clustering (MMC) aims to explore the consistency across different modalities or/and the uniqueness of private information in heterogeneous modalities so as to learn satisfactory clustering result (Xia et al. 2023a; Yang et al. 2022; Hu et al. 2024b; Wang et al. 2020). It has been successfully applied in lots of fields like medical imaging, intelligent transportation systems, and natural language processing. Recently, deep learning has gained widespread attention in many communities due to its efficiency in mining key information from data and its effectiveness in unsupervised learning.

Existing deep MMC methods can be roughly divided into three categories. (1) Graph Neural Network-Based: Capture inter-modal relationships using graph structures. For example, an extended contrastive learning approach is developed to handle out-of-sample data in graph MMC (Xia et al. 2023b). Afterwards, a graph-based structural spectral-spatial clustering framework was proposed (Peng et al. 2023), effectively investigating high-order pixel structure relationships and thereby capturing robust spectral-spatial features and the overarching clustering architecture. (2) Matrix Decomposition-Based: Represent multi-modal data as matrices using techniques like non-negative matrix factorization (NMF). For instance, in the work (Cui et al. 2024), NMF is used to standardize feature scales across views, align modalities, and extract discriminative features. Similarly, sparse constraints with NMF are applied to capture the statistical properties of each modality, thereby enhancing cluster assignments (Li et al. 2024). (3) Joint Representation Learning-Based: Learn a shared latent space for clustering across modalities. For instance, in the work (Yan et al. 2023), a shared representation was obtained by fusing private features of multiple modalities and it was further utilized to guide the model in exploring the relevance of each modality. In the work (Hu et al. 2023), unique feature information between modals was aligned and combined with a triple contrastive learning framework to obtain global features so that the downstream tasks can benefit.

However, the above MMC methods still contain two limitations. First, different aspects of self-guidance information in feature representation and cluster assignment are ignored, leads to unsatisfactory clustering quality. Second, redundant information in heterogeneous multi-modal data hinders the exploration of complementary information between modalities, reducing clustering performance.

To address the above challenges, we propose a novel Multi-aspect Self-guided Deep Information Bottleneck (MSDIB) method for multi-modal clustering, as shown in Figure 1. MSDIB integrates information compression and

*Corresponding authors.

preservation to extract discriminative features from heterogeneous data through mutual guidance between feature and clustering information. The information bottleneck (Hu et al. 2024a) retains features relevant to target Y by creating a compressed representation T for input X, optimized by $\mathcal{L}_{\mathrm{IB}} = I(X;T) - \beta I(T;Y)$, where $I(X;T)$ captures information T retains from X, $I(X;Y)$ its predictive power for Y, and β balances compression and relevance. MSDIB employs three information aspects: private representations for each modality, a shared representation across modalities, and self-supervised pseudo labels. During compression, each modality is reduced to a compact representation guided by others by minimizing mutual information within and between modalities, ensuring uniqueness and discriminability. These private representations are then fused into a global shared representation. In preservation, the global shared representation guides the learning of private representations. Additionally, we maximize the mutual information between the shared clustering assignment and the assignment of each modality, enhancing clustering consistency across modalities. The main contributions are as follows:

- We propose a novel MSDIB method for deep multi-modal clustering, which effectively explores the discriminative information among multi-modal data sources for clustering and can also provide valuable support for other downstream tasks.
- A multi-aspect self-guidance strategy is designed, which employs the private representation of each modality, the shared representation, and the self-supervised pseudo label information of each iteration to self-guide the cluster-friendly representation learning. With the information bottleneck theory as a strong support, the shared information among modalities is retained to the maximum extent, and thus benefiting the clustering performance.
- A unified variation optimization method is designed to fully solve the objective function, which has been demonstrated its effectiveness in various multi-modal datasets.

The Proposed Method

Problem Formulation

Each $\{X^1, X^2, \ldots, X^m\}$ denotes the variable of data input from m modals. The input $\{X^i\}_{i=1}^m$ of each modality contains $\{x_1^i, x_2^i, \ldots, x_n^i\} \in \mathbb{R}^{n \times d^i}$ where n is number of data samples and d^i the feature dimensionality. $\{H^1, H^2, \ldots, H^m\}$ represents the compressed feature representation of input $\{X^i\}_{i=1}^m$. H^s represents the feature representation after fusion of each modality $\{H^i\}_{i=1}^m$. $\{Y^1, Y^2, \ldots, Y^m\}$ represents the local clustering assignment of feature representation $\{H^i\}_{i=1}^m$ obtained through the clustering model, and Y^s represents the global clustering assignment of fused feature H^s. The MSDIB objective is to eliminate redundant multi-modal information through feature compression while preserving useful shared features and self-supervised pseudo labels (predicted labels on unlabeled data). We optimize the multi-aspect self-guided objective using Adam until convergence, resulting in the global cluster assignment C.

Proposed Objective Function

The proposed MSDIB promotes each other in an end-to-end training manner and achieves mutual benefit to achieve our goals. By jointly optimizing the model, satisfactory clustering results are obtained. The overall loss of this model is

$$\mathcal{L}_{total} = \mathcal{L}_{DDC} - \alpha \mathcal{L}_{IP} + \beta \mathcal{L}_{IC}, \tag{1}$$

where $\mathcal{L}_{IP}$ denotes the information preservation part, $\mathcal{L}_{IC}$ denotes the information compression part, $\mathcal{L}_{DDC}$ denotes the clustering module, α and β is the trade-off parameter.

Multi-aspect Self-guidance

The MSDIB method proposed in this paper performs self-guided optimization by utilizing three aspects of guidance information. There are three types of guidance information,

$\{H^i\}_{i=1}^m \Leftrightarrow \{H^j\}_{i=1}^m$, where $i \neq j$ (the mutual guidance between the private representations of each modal). The latent representations of different modal are provided as self-guiding signals for training the latent representations of other modals (Wang et al. 2023). Through this self-guidance signal, MSDIB can effectively explore the complementary information between modalities, thereby enhancing the latent representation learning of other modalities and providing more comprehensive feature representation for downstream clustering tasks.

$H^s \Rightarrow \{H^i\}_{i=1}^m$ (the common representation of each modal guides the private representation of each modal). The shared representation has rich and comprehensive global information, and its guidance from global information can further ensure the consistency of the same object between different modals.

$Y^s \Rightarrow \{Y^i\}_{i=1}^m$ (the clustering assignment of global features guides the clustering assignment of each modal). Global cluster assignment helps similar objects to be grouped into the same cluster and dissimilar objects to be separated, and its guidance would probably help to promote the consistency among local cluster assignments from each modality.

The above self-guided information jointly help the MSDIB method to learn discriminative information and eliminate the redundant information between multiple modalities, and finally ensure the consistency of clustering results. As shown in Figure 1, guiding direction of the cluster space on left is to express similar cluster assignments will be gathered together, and guiding direction on right is that different clusters will be pushed as far as possible. Despite different guiding directions, both clustering spaces follow the same guiding principles.

Information Compression

In the information compression part, we extract the low-dimensional compact feature representations $\{H^i\}_{i=1}^m$ from each modal under a deep encoder network and then mutually compress them to eliminate redundancy. To achieve our goal, we constrain the compression of information between inputs and features by minimizing the mutual information. Therefore, the objective function of this part is formulated

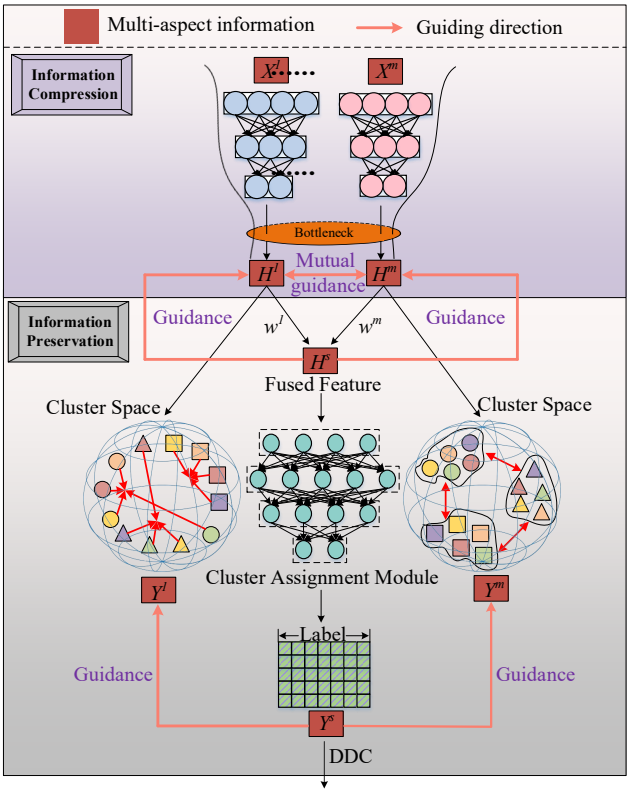

Figure 1: The framework of MSDIB method. The MSDIB processes multi-modal data $\{X^i\}_{i=1}^m$ through a shared-parameter encoder under an information compression mechanism, producing modality-specific feature representations $\{H^i\}_{i=1}^m$. To remove redundant cross-modal information, it minimizes the mutual information $\sum_{i=1}^m I(X^i, H^i)$ and $\sum_{i=1}^m \sum_{j=i+1}^m \mathbb{1}_{i\neq j} I(H^i, H^j)$, allowing the modalities to guide each other's representation learning. Next, a fusion mechanism aggregates the modality-specific features into a shared representation H^s. The local representations $\{H^i\}_{i=1}^m$ inform a clustering module to produce modality-specific assignments $\{Y^i\}_{i=1}^m$, while the shared representation yields a global cluster assignment Y^s. Finally, an information preservation mechanism maximizes $\sum_{i=1}^m I(H^i, H^s)$ and $\sum_{i=1}^m I(Y^i, Y^s)$, capturing discriminative information that enhances clustering performance under the guidance of H^s and Y^s.

by

$$\mathcal{L}_{IC} = \sum_{i=1}^{m} I(X^i, H^i) + \sum_{i=1}^{m} \sum_{j=i+1}^{m} \mathbb{1}_{i\neq j} I(H^i, H^j), \quad (2)$$

where $I(\cdot,\cdot)$ represents the calculation of the mutual information between two variables. If $i \neq j$, the value represented by $\mathbb{1}_{i\neq j}$ is 1, otherwise it is 0. The first term in Eq. (2) calculates the mutual information between the original input and the private information of each modality, with the aim of fully mining the potential features of heterogeneous multi-modal data. The second term is to calculate the mutual information between the private information of each modality. We finally minimize the objective function $\mathcal{L}_{IC}$ to achieve the purpose of eliminating redundant information between each modality and making the private information more complementary under mutual guidance.

Information Preservation

Figure 1 consists of information compression (upper) and information preservation (lower). Compression removes redundant data, maintaining complementary representations. Preservation ensures compressed data is clustering-friendly and consistent across modalities. We fuse private features from multiple modalities using weighted fusion, unlike attention mechanisms that are time-consuming and sensitive to data quality. Instead, we design a simple fusion mechanism that dynamically updates modality-specific weights via backpropagation, similar to parameter optimization. Initially, all weights are equal ($\sum_{i=1}^m w^i = 1$), ensuring balanced early contributions. During training, weights are refined based on the overall objective, effectively integrating diverse modal information.

To obtain modality-specific cluster assignments $\{Y^i\}_{i=1}^m$, we employ a multi-layer perceptron with a softmax output layer, thus mapping features into a cluster space (space for clustering with pseudo labels). Within this space, features are iteratively refined to promote intra-cluster compactness and inter-cluster separability. To maintain generality, the clustering assignment of fusion features is obtained similarly. Therefore, the objective function of this part is

$$\mathcal{L}_{IP} = \sum_{i=1}^{m} I(H^i, H^s) + \sum_{i=1}^{m} I(Y^i, Y^s). \quad (3)$$

The first term in Eq. (3) calculates the mutual information between the private features and the common features. The second term calculates the mutual information between the cluster assignment of each modality and the global cluster assignment, so that the global cluster assignment retains more partitions that are beneficial to the final clustering. By maximizing the objective function $\mathcal{L}_{IP}$, we ensure that the uniqueness of each modality in the multi-modal data is fully utilized, capturing a more comprehensive and reliable cluster assignment.

Deep Divergence-based Clustering Module

To further ensure the compactness of similar cluster assignments and the separability of dissimilar cluster assignments, we introduce a deep divergence clustering (DDC) model. The clustering loss DDC of MSDIB is used for global clustering assignment. The loss comprises three components: (1) Quantifies the relationship between cluster centers and data distributions. (2) Ensures clustering vectors are orthogonal, avoiding correlation and enhancing independence. (3) Enhances model robustness and prevents trivial solutions. By integrating these components, the DDC loss optimizes clustering for multi-modal nonlinear data and accounts for distribution differences across modalities, thereby improving per-

formance. The function of DDC loss is given by

$$\begin{aligned}\mathcal{L}_{DDC} = &\frac{1}{c}\sum_{i=1}^{c-1}\sum_{j>i}\frac{\mu_i^T\mathbf{E}\mu_j}{\sqrt{\mu_i^T\mathbf{E}\mu_i\mu_j^T\mathbf{E}\mu_j}} + triu(C^TC)\\ &+\frac{1}{c}\sum_{i=1}^{c-1}\sum_{j>i}\frac{\gamma_i^T\mathbf{E}\gamma_j}{\sqrt{\gamma_i^T\mathbf{E}\gamma_i\gamma_j^T\mathbf{E}\gamma_j}},\end{aligned} \tag{4}$$

where c is the number of clusters, the term μ_i is the column vector of the clustering result C, and γ_i is determined as the i-th column vector of the matrix $U_{ab} = exp(- \parallel \alpha_a - e_b \parallel^2$), where e_b is the b-th vertex of the simplex. The expression $triu(C^TC)$ represents the sum of the elements of the upper triangular part of the matrix.

Differences with Related Methods

There are mainly two related previous works which utilized the information bottleneck theory to eliminate redundant information and maximize complementary information across various modalities. Unlike SIB-MSC (Wang et al. 2023), we integrated features from all modalities to guide the features of each modality. Unlike DCIB (Yan et al. 2023), we introduced global representation clustering assignments to guide the clustering assignment of each modality. To leverage the rich and comprehensive global information, we utilize this global representation to more effectively extract common information from heterogeneous data, making the clustering results advantageous for downstream tasks.

Optimization

In order to solve our proposed objective function, we designed a variational optimization method to approximate the mutual information into a trainable loss function. Since the mutual information calculation contains two random variables, we have (the 1-th modality as example)

$$\begin{aligned}I(X^1;H^1) &= \int_{h^1}\int_{x^1} p(x^1,h^1)\log\frac{p(x^1,h^1)}{p(x^1)p(h^1)}\\ &= \int_{h^1}\int_{x^1} p(x^1,h^1)\log\frac{p(x^1 \mid h^1)}{p(x^1)},\end{aligned} \tag{5}$$

where $p(x^1, h^1)$ is the joint probability density function of x^1, h^1, and $p(x^1)$, $p(h^1)$ are the marginal probability density functions of x^1 and h^1 respectively.

The difficulty in solving Eq. (5) is that the posterior probability distribution $p(x^1 \mid h^1)$ is unknown and cannot be obtained by calculation. Variational inference can use a parameterized distribution to approximate the target posterior distribution, making the approximation result infinitely close to the posterior distribution. Inspired by variational estimation, we try to use the variational estimate $q(x^1)$ of $p(x^1)$ to approximate the posterior probability distribution $p(x^1 \mid h^1)$. In order to make the distance between $q(x^1)$ and $p(x^1 \mid h^1)$ much closer, we use the Kullback-Leibler (KL) divergence measure to constrain. Since the KL divergence is non-negative, we can get

$$\begin{aligned}&KL(p(x^1)||q(x^1)) = \int p(x^1)\log\frac{p(x^1)}{q(x^1)} > 0\\ &\Rightarrow \int p(x^1)\log p(x^1) > \int p(x^1)\log q(x^1)\\ &\Rightarrow p(x^1) > q(x^1).\end{aligned} \tag{6}$$

Now, $I(X^1; H^1)$ can be rewritten as follows

$$\begin{aligned}I(X^1;H^1) &= \int\int p(x^1,h^1)\log\frac{p(x^1 \mid h^1)}{p(x^1)}\\ &< \int\int p(x^1,h^1)\log\frac{p(x^1 \mid h^1)}{q(x^1)}.\end{aligned} \tag{7}$$

Due to $p(x^1, h^1) = p(h^1)p(x^1|h^1)$, $I(X^i; H^i)$ can be optimized as

$$I(X^i;H^i) < \sum_{i=1}^{m}\int\int p(h^i)p(x^i|h^i)\log\frac{p(x^i \mid h^i)}{q(x^i)}. \tag{8}$$

To eliminate extraneous elements, Monte Carlo sampling (Hastings 1970) is utilized to approximate $p(h^1)$, thereby achieving a more accurate estimation, so we obtain

$$I(X^i;H^i) < \sum_{i=1}^{m}\int p(x^i|h^i)\log\frac{p(x^i \mid h^i)}{q(x^i)}. \tag{9}$$

Suppose $p(x^i|h^i)$ obey a Gaussian distribution, in which its mean μ and variances σ can be learned by the sharing specific encoder (Alemi et al. 2017; Mao et al. 2021). For simplicity, we reparameterize h^i as:

$$h^i = \mu(x^i) + \sigma(x^{i1}) * \theta, \tag{10}$$

where θ denotes standard normal distribution. Next, $I(X^i; H^i)$ can be expressed as

$$\begin{aligned}I(X^i;H^i) &< \sum_{i=1}^{m}\{\mathbb{E}_{\theta_i}\log\frac{p(x^i \mid h^i)}{q(x^i)}\}\\ &< \sum_{i=1}^{m}\mathbb{E}_{\theta_i}\{KL[p(x^i \mid h^i)||q(x^i)]\}.\end{aligned} \tag{11}$$

Moreover, to ensure that the data samples are evenly partitioned into all categories, we set a constraint to $q(x_i)$ based on the uniform distribution as follows:

$$\sum_{i}^{M} q(x_i) = \frac{M}{|C|}, \tag{12}$$

where M is the number of data instances and $|C|$ = K is the number of clusters. Finally,

$$\begin{aligned}I(X^i;H^i) &\approx \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{\theta_i}\{KL[p(x^i \mid h^i)||q(x^i)]\}\\ &\text{with}\sum_{i}^{M} q(x_i) = \frac{M}{|C|}.\end{aligned} \tag{13}$$

Algorithm 1: MSDIB Algorithm

Input: Multi-modal dataset $\{X^i\}_{i=1}^m$; number of clusters k.
Parameter: Hyperparameters α, β, and learning rate γ.
Output: The label predictor C.

1: Initialize the neural network parameters$\{\theta^i\}_{i=1}^m$.
2: **while** not converge **do**
3: Extract modal-specific representations $\{H^i\}_{i=1}^m$ by sharing modal-specific encoders.
4: Compute the IC loss by Eq. (16).
5: Compute the IP loss by Eq. (17).
6: Compute the DDC loss by Eq. (4).
7: Optimize the overall loss Eq. (1) by adam optimizer and back-propagate loss. The neural network parameter update process is as follows:

Update IC loss: $\theta_{IC} \leftarrow \theta_{IC} - \gamma \nabla_{\theta_{IC}} \frac{\partial L_{total}(\theta_{IC})}{\partial \theta_{IC}}$.

Update IP loss: $\theta_{IP} \leftarrow \theta_{IP} - \gamma \nabla_{\theta_{IP}} \frac{\partial L_{total}(\theta_{IP})}{\partial \theta_{IP}}$.

Update DDC loss: $\theta_c \leftarrow \theta_c - \gamma \nabla_{\theta_c} \frac{\partial L_{total}(\theta_c)}{\partial \theta_c}$.

8: **end while**
9: **return** C

Next, we show the optimization of mutual information $I(H^i, H^j)$, $I(H^i, H^s)$ and $I(Y^i, Y^j)$. For $I(H^i, H^j)$, we first calculate the joint probability distribution $p(H^i, H^j)$. Specifically, the initial joint probability is first calculated by expanding the dimension and element-wise multiplication, and then summing over the sample dimensions.

$$p(H^i, H^j) = \sum_{n=1}^{bn} H_n^i \times (H_n^j)^T, \tag{14}$$

where bn represents the batchsize. Then the result is symmetrized to ensure the symmetry of the matrix $p(H^i, H^j) = \frac{1}{2}(p(H^i, H^j) + p(H^j, H^i))$, and finally the joint probability matrix is normalized $p(H^i, H^j) = \frac{P_{ij}}{\sum_{ij} P_{ij}}$. Second, we calculate the marginal probabilities of $p(H^i)$ and $p(H^j)$. Finally, we can calculate $I(H^i, H^j)$ by

$$I(H^i, H^j) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} \mathbb{1}_{i \neq j} p(H^i, H^j) \log \left(\frac{p(H^i, H^j)}{p(H^i) p(H^j)} \right). \tag{15}$$

Similarly, $I(H^i, H^s)$ involves the calculation of mutual information between features and can be optimized similarly to $I(H^i, H^j)$. Likewise, $I(Y^i, Y^j)$ is optimized by calculating the joint probability density between each cluster assignment, then obtaining the joint probability matrix and edge rate similarly to $I(H^i, H^j)$. Therefore, the objective function of information compression and preservation can be transformed into a trainable loss function,

$$\mathcal{L}_{IC} \approx \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{\theta_i} \{KL[p(x^i \mid h^i) || q(x^i)]\} + \frac{1}{M} \sum_{i=1}^{M} \sum_{j=i+1}^{M} I(H^i, H^j). \tag{16}$$

$$\mathcal{L}_{IP} \approx \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{\theta_i} \{KL[p(h^s \mid h^i) || q(h^i)]\} + \frac{1}{M} \sum_{i=1}^{M} I(Y^i, Y^s). \tag{17}$$

To ensure the convergence of the loss function, the final objective function is optimized end-to-end within a unified framework. Algorithm 1 outlines the steps. Parameters θ_{IC}, θ_{IP}, θ_c govern updates for shared encoders, cluster assignment, and clustering model, respectively.

Complexity Analysis

The time complexity of MSDIB splits into three parts: information compression, information preservation, and clustering allocation. Letting D be the input dimension, m the number of modalities, N the sample size, k the number of clusters, and d_c and d_p the maximum IC and IP encoder dimensions, their complexities are: (1) Information compression: $O(mNDd_c)$; (2) Information preservation: $O(N \sum_i^m d_p^2)$; (3) Clustering module: $O(mk^2)$. Overall complexity: $O(K(mNDd_c + N \sum_i^m d_p^2 + mk^2))$, where K is the number of training iterations.

Discussion Between Multi-modal and Uni-modal Clustering Methods Based on Data Augmentations

We discuss them by application scenarios, similarities, and differences: 1) Multi-modal clustering integrates multiple data types (e.g., image-text), while uni-modal clustering focuses on a single type with fewer samples. 2) Both seek discriminative representations to improve clustering. 3) Multi-modal methods emphasize cross-modal fusion and alignment, whereas uni-modal methods rely on data augmentation to increase variability within one modality.

Experiments

Datasets

We evaluate the proposed method by selecting five well-known multi-model datasets, from 1440 to 20,000 in terms of the size of samples.

Caltech-2V (Fei-Fei, Fergus, and Perona 2004) consists of images from 7 categories, totaling 1440 images. It has the feature of Wavelet moments (Shen and Ip 1999) and CENsus TRansform hISTogram (CENTRIST) (Wu and Rehg 2010), where each kind of feature is regarded as a modal.

ESP-Game (Von Ahn and Dabbish 2004) comprises 11,032 images, consisting of 7 categories. The image features and the corresponding text description are used as two modalities.

IAPR (Grubinger et al. 2006) is an image collection with semantic descriptions, consisting of 20,000 images and their corresponding textual descriptions. For this study, a total of 7,855 images with labels no less than 4 were selected and categorized into 6 classes. It utilizes the same two modalities as ESP-Game.

MIRFlickr (Huiskes and Lew 2008) comprises a total of 12,154 images across 6 categories after denoising. It utilizes the same two modalities as ESP-Game.

NUS-Wide (Chua et al. 2009) contains 20,000 images over 8 classes. It comprises a total of two modalities, including both image and text.

Methods	Caltech-2V		IAPR		ESP-Game		MIRFlickr		NUS-Wide	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KM	41.6	30.5	38.9	17.2	48.4	33.5	40.9	22.5	26.8	16.7
Ncuts (TPAMI'00)	39.9	31.2	41.9	18.9	46.5	29.9	48.4	26.1	31.6	14.1
AmKM	46.4	31.4	40.4	17.0	34.9	20.3	41.0	21.6	26.8	15.2
AmNcuts (TPAMI'00)	42.8	5.2	42.2	18.9	33.6	18.9	48.2	26.2	30.4	16.1
CoregMVSC (NIPS'11)	49.2	39.6	35.1	18.4	40.1	28.8	41.0	26.8	26.3	16.8
RMKMC (IJCAI'13)	51.4	33.5	36.4	15.9	44.7	29.7	42.3	23.4	30.5	14.5
SwMC (IJCAI'17)	34.2	26.6	30.2	<u>23.1</u>	43.7	<u>44.2</u>	34.3	<u>34.5</u>	12.5	15.0
ONMSC (AAAI'20)	34.2	26.6	21.6	11.1	17.1	18.1	30.6	16.4	16.9	15.3
TBGL (TPAMI'22)	39.6	34.9	22.6	4.0	22.2	12.0	31.5	28.5	32.1	16.5
MMGC (AAAI'23)	34.9	19.7	26.1	3.3	17.5	1.4	18.9	1.7	40.2	15.9
EAMC (CVPR'20)	40.3	26.6	37.1	16.4	27.1	6.5	30.5	9.1	24.6	9.7
DEMC (INS'21)	37.1	27.9	30.1	13.8	35.5	21.6	44.8	25.2	21.3	11.1
SiMVC (CVPR'21)	51.1	36.9	42.7	18.5	35.3	16.2	45.6	26.3	25.7	10.2
CoMVC (CVPR'21)	59.2	49.2	46.7	21.5	51.8	38.2	49.3	30.6	43.9	31.2
MFLVC (CVPR'22)	61.5	<u>53.6</u>	<u>47.3</u>	22.6	<u>52.1</u>	39.4	<u>53.8</u>	32.8	<u>45.1</u>	<u>33.1</u>
DealMVC (ACM MM'23)	47.6	37.9	35.0	10.8	42.7	24.7	49.3	32.1	36.4	24.4
ICMVC (AAAI'24)	49.6	37.9	37.1	16.8	45.8	29.5	43.5	24.4	34.7	19.8
DIVICE (AAAI'24)	<u>64.1</u>	52.9	45.6	23.0	46.5	27.0	52.3	33.5	36.3	26.0
MSDIB	**67.8**	**55.2**	**50.6**	**26.7**	**66.1**	**46.5**	**58.4**	**37.5**	**48.6**	**33.5**
Free-parameter MSDIB	**65.6**	**54.1**	**50.7**	**25.7**	**65.8**	**46.2**	**55.5**	**35.5**	**47.6**	**33.4**
Ours vs Best Compared	3.7↑	1.6↑	3.3↑	4.1↑	14.1↑	2.3↑	4.6↑	3.0↑	3.5↑	0.4↑

Table 1: Clustering results in terms of ACC and NMI on the multi-model datasets (The values in bold and underline represent the best result and the second best result respectively).

$\mathcal{L}_{DDC}$	$\mathcal{L}_{IC}$	$\mathcal{L}_{IP}$	Caltech-2V		IAPR		ESP-Game		MIRFlickr		NUS-Wide	
			ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
✓			63.4	51.5	47.9	24.8	59.5	43.6	54.4	34.3	45.3	29.9
✓	✓		65.9	52.9	47.7	24.9	59.2	43.4	54.8	34.9	45.2	29.5
✓		✓	65.5	54.1	50.2	25.6	65.1	45.8	56.3	36.0	47.5	33.1
✓	✓	✓	**67.8**	**55.2**	**50.6**	**26.7**	**66.1**	**46.5**	**58.4**	**37.5**	**48.6**	**33.5**

Table 2: Ablation study on the multi-model datasets.

State-of-the-art Methods

We compared the proposed algorithm with various state-of-the-art clustering methods, classified into four types.

Single-modal clustering methods (run the single-modal method on each modal and report the best clustering results): KM (K-Means), Ncuts (Normalized Cuts).

Full-modal clustering methods (connect all models and apply the single-modal clustering method): AmKM (All-modal K-Means), AmNcuts (All-modal NCuts).

Traditional MMC methods: CoregMVSC (Kumar, Rai, and III 2011), RMKMC (Cai, Nie, and Huang 2013), SwMC (Nie, Li, and Li 2017), ONMSC (Zhou et al. 2020), TBGL (Xia et al. 2022), MMGC (Tan et al. 2023).

Deep MMC methods: EAMC (Zhou and Shen 2020), DEMC [1] (Xu et al. 2021), SiMVC and CoMVC [2] (Trosten et al. 2021), MFLVC [3] (Xu et al. 2022), DealMVC [4] (Yang et al. 2023), ICMVC [5] (Chao, Jiang, and Chu 2024) and DIVICE [6] (Lu et al. 2024).

Implementation Details

We implemented the framework in PyTorch 1.13.0 on Windows 10 with a 24 GB NVIDIA RTX-3090 GPU and i7-12700F CPU. Training converged within 100 epochs. We ran the model 20 times, selecting the highest accuracy at the lowest loss to prevent local maxima. The batch size was 100, using Adam with a learning rate of 0.0001. Grid search optimized trade-off parameters α and β in (0, 1) with a step size of 0.1.

For a more comprehensive analysis, the clustering performance was evaluated using two popular metrics: Clustering Accuracy (ACC) and Normalized Mutual Information (NMI). Higher values of these metrics indicate better clustering performance.

[1] https://github.com/SubmissionsIn/DEMVC

[2] https://github.com/DanielTrosten/mvc

[3] https://github.com/SubmissionsIn/MFLVC

[4] https://github.com/xihongyang1999/DealMVC

[5] https://github.com/liunian-Jay/ICMVC

[6] https://github.com/XLearning-SCU/2024-AAAI-DIVIDE

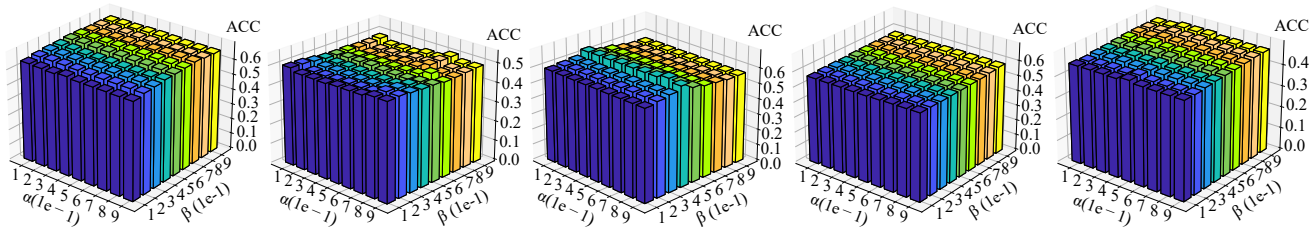

Figure 2: The parameter analysis of the proposed method on all the multi-model datasets with the order of Caltech-2V, IAPR, ESP-Game, MIRFlickr and NUS-Wide.

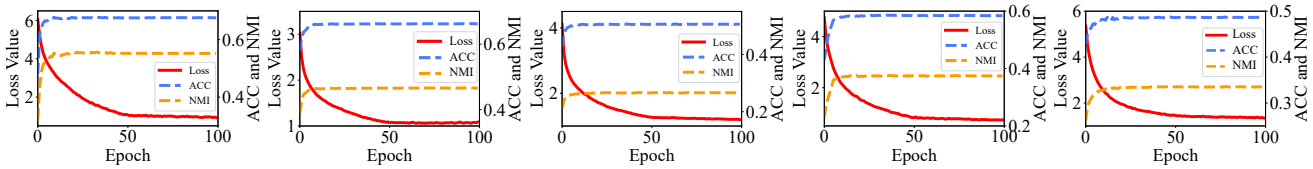

Figure 3: The convergence curves on multi-model datasets.

Clustering Results and Analysis

In this section, we conducted a series of comprehensive experiments to compare the proposed MSDIB with many state-of-the-art clustering methods. Table 1 presents the clustering results on five datasets.

MSDIB outperforms single and full-model clustering by extracting complementary information and eliminating redundant single-modal data. Compared to traditional MMC methods, deep MMC achieves competitive results on most datasets, highlighting the strength of deep learning in discriminative feature representation. Our approach consistently surpasses deep MMC across all datasets, notably improving performance on the ESP-Game dataset. This improvement results from leveraging information bottleneck theory in a self-guided manner for effective information compression and preservation.

Free-parameter MSDIB and MSDIB yield remarkably consistent clustering results, highlighting the stability of the method and low sensitivity to parameter changes. By fixing two parameters to 1, free-parameter MSDIB demonstrates the hyperparameter insensitivity of the method, reducing reliance on complex tuning and simplifying real-world deployment.

Ablation Study

Table 2 validates each MSDIB module's effectiveness. The DDC module alone showed limited success due to its single-modality focus. Nonetheless, the integration of the IC module to compress redundant information within individual modalities and the IP module to preserve the complementary information between modal clusterings signifies a remarkable improvement in clustering performance. The whole optimization of these components ultimately leads to a substantial enhancement in clustering effectiveness.

Parameter Sensitivity

The objective function includes two regularization coefficients for the IP and IC modules. We conducted a sensitivity study across all datasets with various settings, as shown in Figure 2. Using grid search, we tuned trade-off parameters α and β from 0 to 1 in 0.1 increments. Results indicate that ACC values remain consistent across most settings, demonstrating low sensitivity and minimal impact on clustering performance. This phenomenon also highlights the potential of our proposed deep multi-modal clustering approach for more practical applications.

Convergence Analysis

Figure 3 illustrates the convergence curves of the overall loss function, ACC, and NMI metrics across various dataset configurations, specifically including Caltech-2V, ESP-Game, IAPR, MIRFlickr, and NUS-Wide datasets. The graphical representations clearly demonstrate that the convergence of all three metrics stabilizes beyond the 40-epoch threshold. This observation indicates a fast model training process that ensures consistent and stable clustering performance across multiple datasets.

Conclusion

In this paper, we introduce a novel MSDIB method for multi-modal clustering, which extracts discriminative information from heterogeneous data through mutual guidance between feature representations and clustering information. We enhance MSDIB by optimizing three aspects: private feature representations, common feature representations, and self-supervised pseudo labels. Finally, variational optimization ensures the trainability and convergence of the unified objective.

However, MSDIB is designed for general deep information bottleneck multi-modal clustering, it may struggle with incomplete or unaligned data. Missing values cause imbalances and misalignment leads to inconsistent features, limiting its applicability. In the future, we will address the above more complex challenges in practical deep multi-modal clustering problem.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (project no. 62206254 and 62176239) and China Postdoctoral Science Foundation (project no. 2024T170843 and 2023M743186).

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *5th International Conference on Learning Representations, ICLR*.

Cai, X.; Nie, F.; and Huang, H. 2013. Multi-View K-Means Clustering on Big Data. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 2598–2604.

Chao, G.; Jiang, Y.; and Chu, D. 2024. Incomplete contrastive multi-view clustering with high-confidence guiding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11221–11229.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, 1–9.

Cui, G.; Wang, R.; Wu, D.; and Li, Y. 2024. Semi-supervised Multi-view Clustering based on NMF with Fusion Regularization. *ACM Transactions on Knowledge Discovery from Data*, 18(6): 1–26.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, 178–178.

Grubinger, M.; Clough, P.; Müller, H.; and Deselaers, T. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1): 97–109.

Hu, S.; Lou, Z.; Yan, X.; and Ye, Y. 2024a. A Survey on Information Bottleneck. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8): 5325–5344.

Hu, S.; Zhang, C.; Zou, G.; Lou, Z.; and Ye, Y. 2024b. Deep Multiview Clustering by Pseudo-Label Guided Contrastive Learning and Dual Correlation Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Hu, S.; Zou, G.; Zhang, C.; Lou, Z.; Geng, R.; and Ye, Y. 2023. Joint contrastive triple-learning for deep multi-view clustering. *Information Processing & Management*, 60(3): 103284.

Huiskes, M. J.; and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 39–43.

Kumar, A.; Rai, P.; and III, H. D. 2011. Co-regularized Multi-view Spectral Clustering. In *Advances in Neural Information Processing Systems*, 1413–1421.

Li, J.; Kang, P.; Sun, W.; and Jiang, Z. 2024. Local residual preserving non-negative matrix factorization for multi-view clustering. *Neurocomputing*, 600: 128054.

Lu, Y.; Lin, Y.; Yang, M.; Peng, D.; Hu, P.; and Peng, X. 2024. Decoupled Contrastive Multi-View Clustering with High-Order Random Walks. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI*, 14193–14201.

Mao, Y.; Yan, X.; Guo, Q.; and Ye, Y. 2021. Deep Mutual Information Maximin for Cross-Modal Clustering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 8893–8901.

Nie, F.; Li, J.; and Li, X. 2017. Self-weighted Multiview Clustering with Multiple Graphs. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2564–2570.

Peng, B.; Yao, Y.; Lei, J.; Fang, L.; and Huang, Q. 2023. Graph-Based Structural Deep Spectral-Spatial Clustering for Hyperspectral Image. *IEEE Trans. Instrum. Meas.*, 72: 1–12.

Shen, D.; and Ip, H. H. 1999. Discriminative wavelet shape descriptors for recognition of 2-D patterns. *Pattern recognition*, 32(2): 151–165.

Tan, Y.; Liu, Y.; Wu, H.; Lv, J.; and Huang, S. 2023. Metric multi-view graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9962–9970.

Trosten, D. J.; Løkse, S.; Jenssen, R.; and Kampffmeyer, M. 2021. Reconsidering Representation Alignment for Multi-View Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1255–1265.

Von Ahn, L.; and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 319–326.

Wang, Q.; Cheng, J.; Gao, Q.; Zhao, G.; and Jiao, L. 2020. Deep multi-view subspace clustering with unified and discriminative learning. *IEEE Transactions on Multimedia*, 23: 3483–3493.

Wang, S.; Li, C.; Li, Y.; Yuan, Y.; and Wang, G. 2023. Self-supervised information bottleneck for deep multi-view subspace clustering. *IEEE Transactions on Image Processing*, 32: 1555–1567.

Wu, J.; and Rehg, J. M. 2010. Centrist: A visual descriptor for scene categorization. *IEEE transactions on pattern analysis and machine intelligence*, 33(8): 1489–1501.

Xia, W.; Gao, Q.; Wang, Q.; Gao, X.; Ding, C.; and Tao, D. 2022. Tensorized bipartite graph learning for multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 5187–5202.

Xia, W.; Gao, Q.; Wang, Q.; Gao, X.; Ding, C.; and Tao, D. 2023a. Tensorized Bipartite Graph Learning for Multi-View Clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4): 5187–5202.

Xia, W.; Wang, T.; Gao, Q.; Yang, M.; and Gao, X. 2023b. Graph embedding contrastive multi-modal representation

learning for clustering. *IEEE Transactions on Image Processing*, 32: 1170–1183.

Xu, J.; Ren, Y.; Li, G.; Pan, L.; Zhu, C.; and Xu, Z. 2021. Deep embedded multi-view clustering with collaborative training. *Information Sciences*, 573: 279–290.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022. Multi-level Feature Learning for Contrastive Multi-view Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 16030–16039.

Yan, X.; Mao, Y.; Ye, Y.; and Yu, H. 2023. Cross-modal clustering with deep correlated information bottleneck method. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.

Yang, H.; Gao, Q.; Xia, W.; Yang, M.; and Gao, X. 2022. Multiview Spectral Clustering With Bipartite Graph. *IEEE Trans. Image Process.*, 31: 3591–3605.

Yang, X.; Jiaqi, J.; Wang, S.; Liang, K.; Liu, Y.; Wen, Y.; Liu, S.; Zhou, S.; Liu, X.; and Zhu, E. 2023. Dealmvc: Dual contrastive calibration for multi-view clustering. In *Proceedings of the 31st ACM International Conference on Multimedia*, 337–346.

Zhou, R.; and Shen, Y. 2020. End-to-End Adversarial-Attention Network for Multi-Modal Clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14607–14616.

Zhou, S.; Liu, X.; Liu, J.; Guo, X.; Zhao, Y.; Zhu, E.; Zhai, Y.; Yin, J.; and Gao, W. 2020. Multi-View Spectral Clustering with Optimal Neighborhood Laplacian Matrix. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 6965–6972.