ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Engineering Journals
Available on IEEE Xplore®

Research Article
https://doi.org/10.1631/ENG.ITEE.2025.0185

Multi-stage contrastive multi-modal clustering with consistency retained

Yanzheng WANG[§1], Yujun WANG[§1], Fengshuo DAI[2], Xin YANG[1], Xiaoheng JIANG[1], Pei LV[1], Shizhe HU[1✉], Mingliang XU[1,3✉]

[1]*School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China*
[2]*International College of Zhengzhou University, Zhengzhou University, Zhengzhou 450001, China*
[3]*Shandong Bosuan Zhixin Information Technology Co., Ltd., Jinan 250101, China*

Abstract: Contrastive learning has received extensive attention because it can effectively strengthen inter-modal alignment and information complementarity in multi-modal clustering. However, there are two main problems with existing contrastive multi-modal clustering methods: (1) Current methods usually perform contrastive learning at only one or two stages and have imprecise selection strategies for negative sample pairs. (2) Most methods focus only on contrastive learning and ignore the retention of consistency information. To address the above challenges, we propose a framework of multi-stage contrastive multi-modal clustering with consistency retained. First, we design a systematic contrastive learning framework, which includes early–middle–late contrastive learning between single features, fused features, and clusters, respectively, which improves the semantic integrity and information fidelity of the fused representation. Second, we propose a new negative sample pair selection strategy to discriminately select the correct negative sample pair for each sample. Finally, we add a new loss term to late-stage contrastive learning, which considers the consistency of clustering results of the same sample under different modalities for the first time. Therefore, compared to the baseline model, this module has improved the accuracy by an average of 7.4 percentage points (PPs). In addition, we design a consistency retained module to ensure that the model can learn the information of each modality without being disturbed by high/low quality modalities, which improves the accuracy by an average of 2.2 PPs. Experiments conducted on multiple multi-modal datasets demonstrate that the accuracy of our method is 4.1 PPs higher on average compared to the second-best multi-modal clustering method, proving the excellent potential of our method.

Key words: Multi-modal clustering; Contrastive learning; Consistency retention; Negative sample selection

1 Introduction

Contrastive multi-modal clustering (CMC) is an unsupervised clustering method that combines contrastive learning (CL) and multi-modal learning. It aims to mine consistent and discriminative clustering structures from data with multiple modalities. It introduces the mechanism of CL on the basis of multi-modal clustering (MMC), and enhances the semantic alignment and discrimination ability between modalities by constructing positive and negative sample pairs.

MMC has received widespread attention in recent years due to its excellent development prospects in different industries, such as medical and bioinformatics (Herath and Kobti, 2024; Yin et al., 2025), recommendation systems and user behavior analysis (Tong, 2025), Internet of Things (Lin et al., 2018), and machine vision measurement (Pan et al., 2023). CL has stronger feature representation and generalization capabilities (Zhang et al., 2023), making the effect of MMC even

§ Equal contribution
✉ Shizhe HU, ieshizhehu@zzu.edu.cn
✉ Mingliang XU, iexumingliang@zzu.edu.cn
Yanzheng WANG, https://orcid.org/0009-0000-4005-7464
Yujun WANG, https://orcid.org/0009-0007-8684-1300
Fengshuo DAI, https://orcid.org/0009-0007-7509-9315
Xin YANG, https://orcid.org/0009-0008-1130-5914
Xiaoheng JIANG, https://orcid.org/0000-0002-5770-0417
Pei LV, https://orcid.org/0000-0002-2654-0561
Shizhe HU, https://orcid.org/0000-0003-1301-2396
Mingliang XU, https://orcid.org/0000-0002-6885-3451
CLC number: TP391.41

Received: Dec. 23, 2025; Revision accepted: June 10, 2026;
Crosschecked: June 17, 2026; Published online: July 22, 2026

better (Xu et al., 2022).

Existing CMCs can be roughly divided into the following four categories.

1. Early-stage CL: These methods (Tian et al., 2020; Trosten et al., 2021; Lu et al., 2024) perform CL only between single features. For example, Lu et al. (2024) proposed a framework that performs inter- and intra-modal CL in different embedding spaces by randomly walking global data pairs, effectively improving the ability of recognizing positive and negative samples.

2. Middle-stage CL: These methods perform CL only on fusion features. For example, from the perspective of an information bottleneck, Ke et al. (2021) compared and aligned the fusion features with the features of each modality, which can effectively capture the consistency information between modalities.

3. Late-stage CL: These methods (Dai et al., 2022; Yang et al., 2023) perform CL only between single clusters. For example, Dai et al. (2022) proposed a framework that stores feature vectors and calculates contrast loss at the cluster level, which can maintain cluster-level consistency well.

4. Two-stage CL: Current methods (Xu et al., 2022, 2023; Hu et al., 2023; Lou et al., 2025) perform CL in no more than two stages, either at the early and middle stages or the early and late stages. For example, Gu and Zhu (2024) proposed to apply CL during the early and middle stages to align with fusion features, thereby facilitating the learning of shared representations across modalities. Lou et al. (2025) proposed a reliable CL mechanism which ensures that each modality is given a fair weight in CL and prevents learning from unreliable sample pairs during the early and late stages.

Although the above CMC methods have achieved good results, there are still two challenges: (1) Most methods apply CL only at one or two stages, and the selection of positive and negative samples is rough. The fusion feature plays an important role in the final clustering result, and even two-stage CL will ignore the CL of the fusion feature. In addition, the correct selection of negative sample pairs has an important impact on CL, but most current CMCs are relatively rough in the construction of positive and negative sample pairs. They often use different modal representations of the same sample as positive sample pairs, and other samples are all regarded as negative samples. Other methods rely on the intermediate results of the fusion feature in the clustering process to generate pseudo-labels to divide the positive and negative pairs. However, this simple division strategy easily introduces a large number of noise sample pairs, misleading the model to learn the correct features. For example, in a ring dataset, the samples of the inner ring and the outer ring may be close to each other locally, but from the overall structure, they belong to different clusters. If traditional pseudo-label-guided CL is used, it is very likely that the outer ring and the inner ring will be mistakenly classified into the same class because they are too close. (2) Most existing CMCs focus only on CL and ignore the retention of consistency information, which will cause CL to be affected by high/low-quality modalities and fail to learn information from each modality fairly.

To address the above challenges, we propose a new MMC method called multi-stage contrastive multi-modal clustering with consistency retained (MCMC-CR). As shown in Fig. 1,

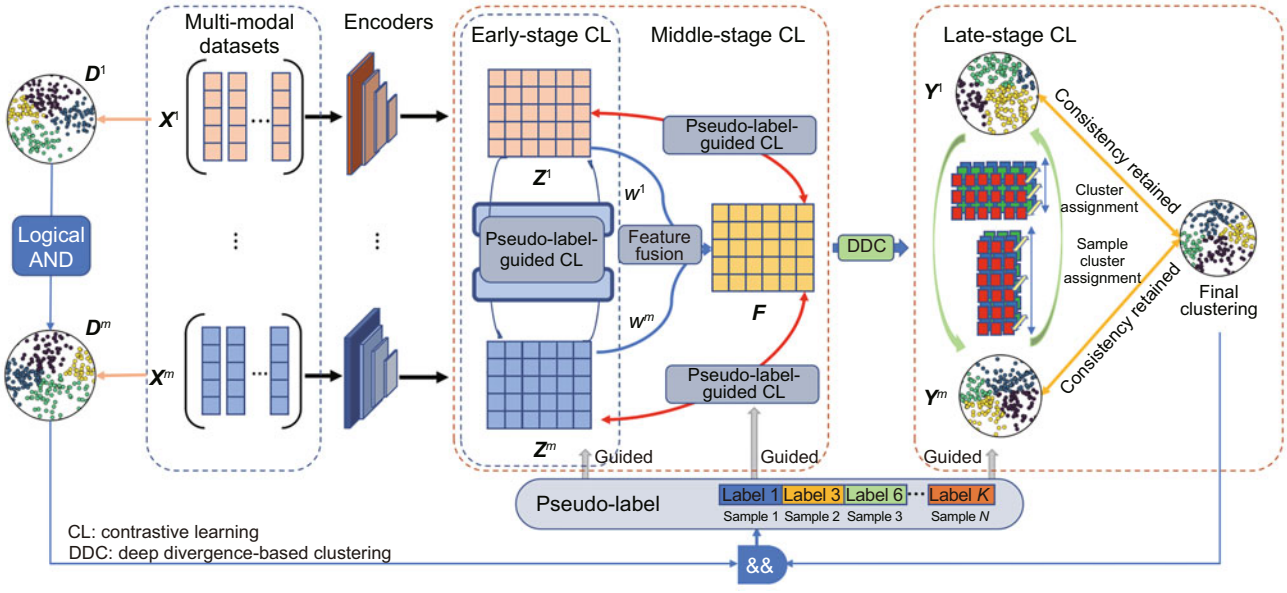

Fig. 1 Framework of the MCMC-CR method. The multi-modal dataset $\{X^m\}_{m=1}^M$ passes through modality-specific encoders to obtain feature $\{Z^m\}_{m=1}^M$. Subsequently, $\{Z^m\}_{m=1}^M$ is adaptively weighted to yield the fused feature F, and is also fed into a clustering module with shared parameters to generate cluster assignment $\{Y^m\}_{m=1}^M$. At the same time, we independently run K-means clustering on the original feature space of each modality to obtain the clustering assignment $\{D^m\}_{m=1}^M$ of samples in the original feature space. CL is applied between single features, between single features and fused features, and between clusters as early, middle, and late stages, respectively. Correct consistency information between modalities can be obtained through CL at multiple stages. $\{D^m\}_{m=1}^M$ and the clustering results C of the fused features jointly generate pseudo-labels to guide all CL. Finally, the consistency information between C and $\{Y^m\}_{m=1}^M$ is calculated to ensure that the final result contains information from each modality

MCMC-CR has three pseudo-label-guided CL stages, namely the early, middle, and late stages, where CL acts on single features, fused features, and clusters. In addition, for all CL, a new negative sample selection strategy is adopted. By considering both the fusion features and the original features of each modality to formulate a new pseudo-label to select the correct positive and negative sample pairs for each sample, false negative sample pairs can be effectively removed. At the same time, we add a consistency retention module to ensure that the model is aligned at the cluster level and reduce the impact of low-quality modalities on clustering. In addition, we add sample-level consistency loss in later-stage CL to ensure the semantic similarity of clustering results of the same sample in different modalities. In summary, the main contributions can be summarized as follows:

1. We propose a new multi-stage CL MMC framework. This method applies CL at early–middle–late stages to ensure global alignment of consistency information between modalities.

2. We design a new negative sample selection strategy that can more accurately construct positive and negative sample pairs. By effectively identifying and eliminating most potential false negative sample pairs, the sample pair quality and clustering robustness in CL are significantly improved.

3. We add a new consistency learning constraint for sample cluster assignment in late-stage CL, which enables the model to better capture the overall structure of the data and prevent the semantics of samples in the same cluster assignment from shifting. This method has shown excellent performance on multiple multi-modal datasets.

2 The proposed method

2.1 Problem formulation

Assume that a multi-modal dataset is $\{\boldsymbol{X}^1, \boldsymbol{X}^2, \ldots, \boldsymbol{X}^M\}$, which is a set of data variables from m (m=1, 2, …, M) modalities. The input $\{\boldsymbol{X}^m\}_{m=1}^M$ of each modality contains $\{\boldsymbol{X}_1^m, \boldsymbol{X}_2^m, \ldots, \boldsymbol{X}_N^m\} \in \mathbb{R}^{N \times d_m}$, where N represents the number of samples and d_m represents the dimension of the feature. The goal of MCMC-CR is to maximize the retention of consistency information between modalities and finally obtain the final clustering assignment $\boldsymbol{C}$ through a clustering method based on deep divergence.

2.2 Overall framework

The proposed method is optimized by minimizing the following loss function:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{MCL}} + \beta \mathcal{L}_{\text{Consistency}} + \mathcal{L}_{\text{DDC}}, \tag{1}$$

where $\mathcal{L}_{\text{MCL}}$ denotes the multi-stage CL loss, which integrates the CL losses from the early, middle, and late stages to ensure global alignment of consistency information between modalities. $\mathcal{L}_{\text{Consistency}}$ is the consistency learning loss, which is used to align the clustering assignments of each modality with the clustering assignments of the fusion features. $\mathcal{L}_{\text{DDC}}$ is the deep divergence-based cluster (DDC) loss. α and β are trade-off hyperparameters that regularize the contributions of multi-stage CL loss and consistency learning loss. Their search ranges are set as $\alpha, \beta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$ for grid search in experiments, and the empirically determined optimal values are $\alpha = 1$ and $\beta = 0.1$.

2.3 Negative sample pair selection module

The correct selection of negative sample pairs plays a crucial role in CMCs. However, in the current selection of positive and negative samples in CL, either instances of different modalities of the same sample as positive sample pairs and the remaining instances as negative sample pairs, or only pseudo-labels generated during the algorithm execution are used to distinguish positive and negative sample pairs. The former may cause samples within the same cluster to be separated, while the latter is highly dependent on the prediction ability of the initial model, and incorrect pseudo-labels may be continuously reinforced by the model, thereby increasing the possibility of model prediction errors.

To address the aforementioned issues, a straightforward method for filtering false negative pairs is to exploit the distance between samples in the original feature space (Cui et al., 2024). This approach is viable because samples assigned to the same cluster are expected to exhibit proximity in the original feature space. Motivated by this, we employ the K-means clustering algorithm (Hartigan and Wong, 1979) on the original feature space of each modality to determine the intra-modal clustering structure of samples within their respective original feature spaces. Based on the clustering result $\boldsymbol{D}^m$ for each modality, the negative sample pair relation matrix $\boldsymbol{D}$ is obtained using the following formula:

$$D_{i,j} = \begin{cases} 1, & \text{if } \bigwedge_{m=1}^{M} \left(D_{i,j}^m = 1 \vee D_{j,i}^m = 1 \right), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $D_{i,j}^m = 1$ indicates that samples i and j are assigned to the same cluster by K-means in the original feature space of the m^{th} modality, and $D_{i,j}^m = 0$ indicates that they belong to different clusters. The logical conjunction $\bigwedge_{m=1}^{M}$ represents a joint decision across all modalities: $D_{i,j} = 1$ if the sample pair (i, j) is clustered into the same category in all M modalities' original feature spaces; otherwise, $D_{i,j} = 0$. Clearly, matrix $\boldsymbol{D}$ is symmetric.

To achieve a more accurate representation of the relationships between sample cluster assignments, we attempt to iteratively update the negative sample pair matrix by integrating pseudo-label information derived from fused feature clustering with the $\boldsymbol{D}^m$ matrix in each iteration. Specifically, in our model, erroneous negative sample pairs are defined as those sample pairs that share the same clustering assignments in the original feature space and the fused feature space, which is formulated as

$$\mathcal{N}_i^{\text{false}} = \{j \mid \mathcal{D}_{i,j} = 1,\ \text{pred}_i = \text{pred}_j,\ j \in [1, N]\}, \tag{3}$$

where $\mathcal{D}_{i,j} = 1$ indicates that samples i and j have the same pseudo-label in the original feature space of all modes, and $\text{pred}_i = \text{pred}_j$ indicates that the fused features of samples i and sample j have the same clustering assignment. N is the number of samples.

Furthermore, we define matrix $\boldsymbol{\mathcal{M}} \in \{0, 1\}^{N \times N}$ as the mask matrix, where each element $\mathcal{M}_{ij}$ indicates whether the

j^{th} sample is an erroneous negative sample for the i^{th} sample. Specifically, each row of the matrix is denoted by $\mathcal{N}_i^{\text{false}}$, representing the erroneous negative samples associated with the i^{th} sample.

2.4 Multi-stage CL module

2.4.1 Early-stage CL

In early CL, a multi-modal dataset $\{\boldsymbol{X}^m\}_{m=1}^M$ passes through modality-specific encoders to generate a feature set $\{\boldsymbol{Z}^m\}_{m=1}^M$. CL in the feature set $\{\boldsymbol{Z}^m\}_{m=1}^M$ can align with the semantic representations of different modalities for the same sample and reduce the representation bias between modalities, thereby strengthening multi-modal collaborative understanding and improving consistency between modalities. For the u^{th} and v^{th} modalities, the CL loss $\ell_{\text{fea}}^{u,v}$ between them can be defined as

$$\ell_{\text{fea}}^{u,v} = -\frac{1}{N}\sum_{i=1}^{N} \ln \frac{\mathrm{e}^{s\left(\boldsymbol{z}_i^u, \boldsymbol{z}_i^v\right)/\tau_1}}{\sum_{s' \in \text{Neg}\left(\boldsymbol{z}_i^u, \boldsymbol{z}_i^v\right)} \mathrm{e}^{s'/\tau_1}}, \tag{4}$$

with

$$s_{ii}^{u,v}\left(\boldsymbol{z}_i^u, \boldsymbol{z}_i^v\right) = \frac{\left(\boldsymbol{z}_i^u\right)^{\mathrm{T}} \boldsymbol{z}_i^v}{\|\boldsymbol{z}_i^u\| \cdot \|\boldsymbol{z}_i^v\|}.$$

τ_1 is a temperature hyperparameter, and $\boldsymbol{z}_i^u$ and $\boldsymbol{z}_i^v$ represent the i^{th} sample from $\boldsymbol{Z}^u$ and $\boldsymbol{Z}^v$, respectively. $s_{ii}^{u,v}$ represents the cosine distance between two samples. $\text{Neg}\left(\boldsymbol{z}_i^u, \boldsymbol{z}_i^v\right)$ represents the set of negative sample pairs, and we select the sample when $\mathcal{M}_{ij} \neq 1$ is the case, because $\mathcal{M}_{ij} = 1$ means that j is a wrong negative sample of i. These false negative pairs are eliminated in a self-supervised manner (Hu et al., 2025a) to avoid compromising the compactness of the clusters. Therefore, following common practice, the early CL loss $\mathcal{L}_{\text{Early}}$ can be expressed as

$$\mathcal{L}_{\text{Early}} = \frac{1}{2}\sum_{u=1}^{M}\sum_{v=1, v\neq u}^{M} \ell_{\text{fea}}^{u,v}. \tag{5}$$

2.4.2 Middle-stage CL

To fully learn more of the information between the modalities, we obtain the fused feature representation by weighted averaging of the feature $\boldsymbol{F}$ of each modality:

$$\boldsymbol{F} = \sum_{m=1}^{M} w^m \boldsymbol{Z}^m, \tag{6}$$

where w^m represents the weight of each modality in the fusion process, satisfying $w^m > 0$ and $\sum_{m=1}^{M} w^m = 1$. These fusion weights are learnable during training. To optimize these weights easily, we select a set of unnormalized parameters as network parameters and train them in the optimization loss function. Finally, these parameters pass through the softmax function to obtain weights. Most existing CMCs only have early-stage CL, and ignore middle-stage CL of the fusion feature. Without middle-stage CL, the fusion features may be biased towards the modality with more information and fail to accurately capture the key information of each modality. As a result, using the fusion features for the final clustering may lead to performance degradation. The middle-stage CL loss $\mathcal{L}_{\text{Middle}}$ can be expressed as follows:

$$\mathcal{L}_{\text{Middle}} = \frac{1}{2}\left(\sum_{m=1}^{M} \ell_{\text{fea}}^{f,m} + \sum_{m=1}^{M} \ell_{\text{fea}}^{m,f}\right), \tag{7}$$

where $\ell_{\text{fea}}^{f,m}$ and $\ell_{\text{fea}}^{m,f}$ denote the contrastive losses from the fusion feature f to the m^{th} modality's feature and from the m^{th} modality's feature to f, respectively. Middle-stage CL uses the same method of eliminating false negative pairs as in early-stage CL. Incorporating contrastive loss between individual modality features and the fused representation encourages the model to preserve the salient information from each modality within the joint representation.

2.4.3 Late-stage CL

Late-stage CL compares the clustering results of each modality to capture higher-level semantic distribution and category information to enhance semantic consistency. Specifically, the feature vector of each modality is clustered through the clustering module to obtain the cluster assignment $\{\boldsymbol{Y}^m\}_{m=1}^M$, with its column $\boldsymbol{y}_i^m$ representing the cluster assignment of the i^{th} sample in the m^{th} modality. Through Eq. (4), we can obtain the cosine similarity of the two cluster assignments, denoted as $s\left(\boldsymbol{y}_i^u, \boldsymbol{y}_j^v\right)$. The late-stage CL for the u^{th} and v^{th} modalities can be expressed as

$$\ell_{\text{clu}}^{u,v} = -\frac{1}{K}\sum_{i=1}^{K} \ln \frac{\mathrm{e}^{s(\boldsymbol{y}_i^u, \boldsymbol{y}_i^v)/\tau_2}}{\mathcal{S}_i^{u,v}}, \tag{8}$$

where

$$\mathcal{S}_i^{u,v} = \sum_{j=1, j\neq i}^{K} \mathrm{e}^{s(\boldsymbol{y}_i^u, \boldsymbol{y}_j^u)/\tau_2} + \sum_{j=1}^{K} \mathrm{e}^{s(\boldsymbol{y}_i^u, \boldsymbol{y}_j^v)/\tau_2}, \tag{9}$$

K denotes the number of clusters, and τ_2 denotes the temperature hyperparameter. The cluster assignment consistency contrast loss can be expressed as

$$\mathcal{L}_{\text{Late\_clu}} = \frac{1}{2}\sum_{u=1}^{m}\sum_{v=1, v\neq u}^{m} \ell_{\text{clu}}^{u,v} - \sum_{u=1}^{m} H(\boldsymbol{Y}^u), \tag{10}$$

where $H(\boldsymbol{Y}^u) = -\sum_{i=1}^{K}[P(\boldsymbol{y}_i^u)\ln P(\boldsymbol{y}_i^u)]$ is the entropy of cluster assignment probabilities $P(\boldsymbol{y}_i^u) = \frac{1}{N}\sum_{t=1}^{N} y_{ti}^u$, and $\sum_{u=1}^{m} H(\boldsymbol{Y}^u)$ is a regularization term that prevents the model from collapsing. $\frac{1}{2}\sum_{u=1}^{m}\sum_{v=1, v\neq u}^{m} \ell_{\text{clu}}^{u,v}$ is designed to learn the consistency of clustering results for individual samples across modalities.

Almost all CMCs use only Eq. (10), which results in focusing only on solving the consistency of clustering assignments between modalities, while ignoring the consistency of clustering results of single samples under different modalities. In reality, for the same data sample, the clustering results in different modalities should be semantically similar. In our model, we fully exploit both types of correlation information to optimize the learning of feature representations. Specifically, the clustering module also generates a sample cluster assignment $\{\boldsymbol{R}^m\}_{m=1}^M$, with its column $\boldsymbol{r}_i^m$ representing the probability that the i^{th} sample is assigned to a different cluster in the m^{th} modality. In particular, $\{\boldsymbol{R}^m\}_{m=1}^M$ is the transpose of $\{\boldsymbol{Y}^m\}_{m=1}^M$. The expression of sample clustering contrast loss

for the two modalities u and v is as follows:

$$\ell_{\text{sample}}^{u,v} = -\frac{1}{N}\ln\sum_{i=1}^{N}\frac{\mathrm{e}^{s(\boldsymbol{r}_i^u,\boldsymbol{r}_i^v)/\tau_1}}{\mathcal{P}_i^{u,v}}, \tag{11}$$

where

$$\mathcal{P}_i^{u,v} = \sum_{\substack{j=1\\ \mathcal{M}_{ij}\neq 1}}^{N}\left(\mathrm{e}^{s(\boldsymbol{r}_i^u,\boldsymbol{r}_j^u)/\tau_1} + \mathrm{e}^{s(\boldsymbol{r}_i^u,\boldsymbol{r}_j^v)/\tau_1}\right). \tag{12}$$

The same negative sample selection strategy is adopted. The sample cluster CL loss can be calculated as follows:

$$\mathcal{L}_{\text{Late\_sample}} = \frac{1}{2}\sum_{u=1}^{M}\sum_{v=1,v\neq u}^{M}\ell_{\text{sample}}^{u,v}. \tag{13}$$

By learning the consistency of sample clustering assignments, the model can more accurately capture the overall structure of the data while avoiding the deviation of semantic labels of samples in the same cluster. The late-stage CL loss can be expressed as

$$\mathcal{L}_{\text{Late}} = \mathcal{L}_{\text{Late\_clu}} + \mathcal{L}_{\text{Late\_sample}}. \tag{14}$$

In summary, the overall loss function of the multi-stage CL module can be expressed as follows:

$$\mathcal{L}_{\text{MCL}} = \mathcal{L}_{\text{Early}} + \mathcal{L}_{\text{Middle}} + \mathcal{L}_{\text{Late}}. \tag{15}$$

2.5 Consistency learning module

As shown by Xu et al. (2023), high/low-quality modalities naturally exist in multi-modal datasets and can degrade CL. To reduce the impact of low-quality modalities on clustering effects at the cluster level and enhance the reliability of semantic labels of specific modalities, we design a consistency learning module. Specifically, by minimizing the Kullback–Leibler (KL) divergence (Joyce, 2011) between the cluster assignment of the fused features and the cluster assignment of each modality, we can not only align them at the cluster level, but also supplement the clustering information that may be lost in each modality. The calculation formula is as follows:

$$\mathcal{L}_{\text{Consistency}} = \sum_{m=1}^{M}\mathrm{KL}\left(\boldsymbol{C}_{\text{f}}\|\boldsymbol{Y}^m\right) = \sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{m=1}^{M}c_{ij}\ln\frac{c_{ij}}{y_{ij}^m}, \tag{16}$$

where $\boldsymbol{C}_{\text{f}}$ is the cluster assignment obtained by fusion features, c_{ij} represents the probability that the i^{th} sample belongs to the j^{th} cluster, and y_{ij}^m represents the probability that the i^{th} sample in the m^{th} modality belongs to the j^{th} cluster.

2.6 Clustering module

To ensure the accuracy of the selection of positive and negative sample pairs during the iteration process, we adopt an effective loss function for clustering. DDC (Kampffmeyer et al., 2019) is a deep learning-based clustering approach that aims to simultaneously learn data representations and uncover clustering structures in unlabeled data by optimizing a discriminative loss function based on information-theoretic divergence metrics. The overall function of DDC loss is given by

$$\begin{aligned}\mathcal{L}_{\text{DDC}} =& \frac{1}{c}\sum_{i=1}^{c-1}\sum_{j>i}\frac{\boldsymbol{\mu}_i^{\text{T}}\boldsymbol{E}\boldsymbol{\mu}_j}{\sqrt{\boldsymbol{\mu}_i^{\text{T}}\boldsymbol{E}\boldsymbol{\mu}_i\boldsymbol{\mu}_j^{\text{T}}\boldsymbol{E}\boldsymbol{\mu}_j}} + \mathrm{triu}(\boldsymbol{G}^{\text{T}}\boldsymbol{G})\\ &+ \frac{1}{c}\sum_{i=1}^{c-1}\sum_{j>i}\frac{\boldsymbol{\gamma}_i^{\text{T}}\boldsymbol{E}\boldsymbol{\gamma}_j}{\sqrt{\boldsymbol{\gamma}_i^{\text{T}}\boldsymbol{E}\boldsymbol{\gamma}_i\boldsymbol{\gamma}_j^{\text{T}}\boldsymbol{E}\boldsymbol{\gamma}_j}}.\end{aligned} \tag{17}$$

The equation consists of three main components. The first term is a generalized version of the Cauchy–Schwarz divergence, which maximizes the divergence between clusters to achieve cluster separation and compactness. Here, $\boldsymbol{E}$ indicates the Gaussian kernel-based matrix, c is the total number of clusters, and $\boldsymbol{\mu}_i$ indicates the i^{th} column of the obtained clustering result $\boldsymbol{G}$. The second term ensures that the vectors of clustering results are orthogonal. The expression $\mathrm{triu}(\boldsymbol{G}^{\text{T}}\boldsymbol{G})$ denotes the sum of the elements in the upper triangular portion of matrix $\boldsymbol{G}$. The third term enforces the clustering assignments to approach the simplex corners by applying the geometric structure to the Cauchy–Schwarz divergence, thereby avoiding trivial solutions. Here, $\boldsymbol{\gamma}_i$ denotes the i^{th} column of matrix $\boldsymbol{V}$, with $[\boldsymbol{V}_{ab}] = \exp(-\|\boldsymbol{\alpha}_a - \boldsymbol{e}_b\|^2)$, where $\boldsymbol{\alpha}_a$ and $\boldsymbol{e}_b$ are the soft clustering assignment vector of the a^{th} sample and the b^{th} one-hot standard simplex vertex vector, respectively.

2.7 Optimization

The details of MCMC-CR are shown in Algorithm 1. Specifically, the model consists of multiple encoders and multiple multi-layer perceptron (MLP)layers, and a common mini-batch gradient descent algorithm is used to optimize the model (Kingma and Ba, 2017).

Algorithm 1 MCMC-CR algorithm

Input: multi-modal dataset $\{\boldsymbol{X}^m\}_{m=1}^M$; hyperparameters τ_1 and τ_2; number of clusters K
Output: final cluster assignment $\boldsymbol{C}$

```
initialize the neural network
while the loss function Eq. (1) does not converge do
    extract the feature representation {Z^m}_{m=1}^M of each modality through modality-specific encoders, and obtain the cluster assignment {Y^m}_{m=1}^M
    obtain the fusion feature through Eq. (6)
    compute early-stage CL loss by Eq. (5)
    compute middle-stage CL loss by Eq. (7)
    compute late-stage CL loss by Eq. (14)
    calculate consistency learning loss by Eq. (16)
    calculate DDC loss by Eq. (17)
    update all network parameters by Eq. (1)
end while
return C
```

2.8 Complexity analysis

In this subsection, we analyze the time complexity of the proposed method. To simplify the complexity analysis, we make the following assumptions: The dimensionalities of the original multi-modal samples are assumed to be identical, denoted by D; The dimensionalities of each layer in the MLP network are also assumed to be the same, denoted by P_{d}. Furthermore, we denote the number of layers in the MLP

network as M_{L}, the number of samples as N_{s}, the number of views as V, the maximum number of neurons in the hidden layer of the encoder as Z, and the number of clusters as C. Based on the above definition, we give the time complexity of optimizing each module as follows. First, we perform K-means clustering in the original feature space, and the time complexity of this step is $\mathcal{T}_{K\text{-means}} = O(CDN_{\mathrm{s}}V)$. $\mathcal{L}_{\mathrm{Early}}$, $\mathcal{L}_{\mathrm{Middle}}$, and $\mathcal{L}_{\mathrm{Late}}$ are all computationally expensive modules. Specifically, the time cost of optimizing the $\mathcal{L}_{\mathrm{Early}}$ module is $\mathcal{T}_{\mathrm{Early}} = O(VN_{\mathrm{s}}DP_{\mathrm{d}} + VN_{\mathrm{s}}(M_{\mathrm{L}}-1)P_{\mathrm{d}}^2 + V^2)$. The time complexity of the self-attention fusion mechanism is $O(N_{\mathrm{s}}VZ^2)$, and the complexity of calculating the CL loss between the modal-specific and fusion features is $O(V^2)$. Accordingly, the time complexity of $\mathcal{L}_{\mathrm{Middle}}$ is $\mathcal{T}_{\mathrm{Middle}} = O(N_{\mathrm{s}}VZ^2 + V^2)$. For cluster assignment consistency CL, the time complexity is $O(CN_{\mathrm{s}}(N_{\mathrm{s}}-1) + C(C-1)N_{\mathrm{s}}^2 + V^2) \approx O(N_{\mathrm{s}}^2C^2 + V^2)$. Sample clustering assignment consistency CL actually performs similar operations, so the time complexity of the $\mathcal{L}_{\mathrm{Late}}$ module is $\mathcal{T}_{\mathrm{Late}} = O(N_{\mathrm{s}}^2C^2 + V^2)$. The time complexity for the $\mathcal{L}_{\mathrm{Consistency}}$ module is $\mathcal{T}_{\mathrm{Consistency}} = O(N_{\mathrm{s}}VC)$. Finally, the time complexity for $\mathcal{L}_{\mathrm{DDC}}$ is $\mathcal{T}_{\mathrm{DDC}} = O(VC^2)$. In summary, the overall time complexity of the proposed method is $\mathcal{T}_{\mathrm{total}} = \mathcal{T}_{K\text{-means}} + K_{\mathrm{t}}(\mathcal{T}_{\mathrm{Early}} + \mathcal{T}_{\mathrm{Middle}} + \mathcal{T}_{\mathrm{Late}} + \mathcal{T}_{\mathrm{Consistency}})$, where K_{t} denotes the number of training iterations.

3 Experiments

3.1 Datasets

We evaluate the effectiveness of our proposed method using five well-known datasets: Caltech-3V, Caltech-4V, NUS-Wide, Flickr, and IAPR. IAPR is short for the International Association for Pattern Recognition Benchmark Dataset. A brief overview of these datasets is presented in Table 1. The Caltech image dataset (Li et al., 2004), consisting of 1440 samples spread over seven unique categories, is accessible in two different multi-modal configurations. Caltech-3V incorporates features of wavelet moments (Shen and Ip, 1999), census transform histogram (Wu JX and Rehg, 2011), and local binary pattern (Wolf et al., 2008), where each type of feature is treated as a separate modality. Caltech-4V extends Caltech-3V by adding generalized search trees (Friedman, 1979) as a new feature modality. NUS-Wide (Chua et al., 2009) is a large-scale image dataset containing 20 000 samples across eight categories, with each sample described by four modalities. Flickr (Huiskes and Lew, 2008) is a widely used dataset in the field of multi-modal learning, containing six categories with 12 154 samples and three features. IAPR (Grubinger et al., 2006) consists of 7855 images spanning six categories, with each sample represented by two modalities: image features and corresponding textual descriptions.

Table 1 Details about the multi-modal datasets

Dataset	Number of samples	Number of categories	Dimension
Caltech-3V	1440	7	40\254\928
Caltech-4V	1440	7	40\254\928\512
NUS-Wide	20 000	8	100\100\100\100
Flickr	12 154	6	100\100\100
IAPR	7855	6	100\100

\ is used to distinguish between the features of different modalities within a dataset

3.2 Comparison methods

To better demonstrate the advantages of our proposed method, we perform comparative experiments with a diverse range of 19 baseline approaches, categorized into three distinct groups.

1. Single-modal clustering methods: Four single-modal clustering methods are compared, i.e., K-means (KM), normalized cuts (Ncuts) (Shi and Malik, 2000), all-modal K-means (AmKM), and all-modal NCuts (AmNcuts). The latter two methods first concatenate all modal features, and then adopt the aforementioned single-modal clustering methods.

2. Traditional MMC methods: Five traditional MMC approaches are included in the comparison, where the details are as follows. CoregMVSC (Kumar et al., 2011) is a co-regularized multi-view spectral clustering method by imposing centroid-based co-regularization on view-specific embeddings and cross-view consistent clustering learning. RMKMC (Cai et al., 2013) is a novel large-scale multiview K-means clustering method with enhanced robustness guarantees. SwMC (Nie et al., 2017) is a self-weighted multiview clustering method by leveraging the Laplacian rank constraint to fuse multiple graphs adaptively and eliminate postprocessing dependency. ONMSC (Zhou SH et al., 2020) is a spectral clustering approach that integrates the optimal neighborhood to enhance representation capability and capture high-order structural information. SMCMB (Lao et al., 2024) is a scalable multi-view clustering approach that generates diversified anchor sets for each view, jointly learns many bipartite graphs via anchor-based subspace representation, and fuses base clusterings from these graphs into a unified bipartite graph for final clustering.

3. Deep MMC methods: Ten deep MMCs are incorporated into the comparative experiments. EAMC (Zhou RW and Shen, 2020) is an MMC method that leverages adversarial learning and attention mechanisms to align latent feature distributions and quantify modal characteristics. DEMC (Xu et al., 2021) is a deep embedded multi-view clustering method that integrates a collaborative training strategy to jointly optimize feature representation learning and clustering structure refinement. SiMVC (Trosten et al., 2021) is a simple deep MMC method without integrating cross-view contrastive alignment mechanisms. CoMVC (Trosten et al., 2021) is an extension of the SiMVC method with the integration of a lightweight multi-modal alignment module. MFLVC (Xu et al., 2022) is a deep MMC method that adopts fusion-free multi-level feature learning and CL to ease objective conflicts. SEM (Xu et al., 2023) is an MMC method that uses self-weighting and reconstruction regularization to mitigate representation degeneration in CL. DIVIDE (Lu et al., 2024) is a method that conducts inter- and intra-view CL across distinct embedding spaces, while employing random walks to distinguish intra-neighborhood negatives from extra-neighborhood positives. SCMVC (Wu S et al., 2024) is a deep MVC method that adopts hierarchical fusion and self-weighted CL, separating objectives to mitigate representation degeneration. SSLN-MVC (Yan et al., 2025) is a self-supervised semantic soft label learning multi-view clustering approach that learns consensus

high-level features and calibrates semantic labels via CL and KL divergence. MSDIB (Hu et al., 2025b) is a multi-aspect self-guided deep information bottleneck for multi-view clustering approach, and extracts discriminative information and eliminates redundancy via compression and preservation.

3.3 Evaluation metrics

The clustering performance is meticulously assessed by employing two widely recognized evaluation metrics: accuracy (ACC) and normalized mutual information (NMI). ACC quantifies the consistency between the predicted cluster labels and the ground-truth annotations, whereas NMI evaluates the information-theoretic overlap between the obtained clusters and the true label distribution.

3.4 Implementation details

In our experiments, we observe that 100 training epochs are sufficient for the algorithm to converge. Consequently, we set the number of training epochs to 100 and run each experiment independently 10 times. To mitigate output biases caused by the model getting trapped in local optima during training, we report the clustering results corresponding to the minimum clustering loss. We set $\alpha = 1$ and $\beta = 0.1$ to scale the magnitudes of the loss functions for multi-stage CL and consistency learning, respectively. The temperature parameters τ_1 and τ_2 for these two processes are empirically set to 0.5 and 1.0, respectively. These parameters are fixed in subsequent experiments to simplify the model training process while ensuring the reproducibility of the experimental results. For all datasets, the training batch size is uniformly set to 128, and we use the Adam optimizer with an initial learning rate of 0.0003.

3.5 Clustering results and analysis

In this subsection, we conduct extensive and rigorous experiments on five publicly accessible datasets to assess the performance of our proposed method in comparison with a suite of state-of-the-art clustering approaches. As shown in Table 2, our method achieves superior clustering performance compared to single-modality and all-modality approaches, as well as traditional and deep MMC methods. The remarkable results demonstrate the strong clustering capability of our method, which is attributed to pseudo-label-guided multi-stage CL combined with consistency learning.

The proposed method demonstrates exceptional multi-modal integration capabilities and robust generalization capabilities. On the Caltech dataset, ACC and NMI of the model exhibit a steady upward trend as the number of modalities increases. This improvement stems from the rich complementary information inherent in multi-modal data. With more modalities available, the model can extract data features from additional dimensions. These results indicate that the proposed method effectively integrates information from diverse modalities while fully leveraging the distinctive features and clustering information present within each modality, thereby significantly enhancing clustering accuracy and consistency. Specifically, on the Caltech-3V dataset, our method shows significant gains over the second-best method (SCMVC), improving ACC and NMI by 4.7 percentage points (PPs) and 3.6 PPs, respectively. Extended to Caltech-4V, it achieves 87.5% ACC and 78.3% NMI, outperforming SCMVC by 3.1 PPs and 5.4 PPs, respectively. Our method demonstrates competitive performance on large-scale datasets. For example, on NUS-Wide, compared to SCMVC, MFLVC, and CoMVC, MCMC-CR surpasses their results by margins of 5.1, 8.8, and 10.0 PPs for ACC, respectively. Across all five multi-modal datasets evaluated, our method achieves significant performance gains, with average improvements of 4.1 PPs in ACC and 3.2 PPs in NMI over the second-best comparison method. These consistent and substantial gains demonstrate the robustness and stability of our approach.

3.6 Ablation study

In this subsection, we conduct an ablation study to highlight the efficacy of each module and to emphasize the importance of the newly designed pseudo-label-guided mechanism.

3.6.1 Effectiveness of each module

To thoroughly analyze the contribution of each component in our proposed method, we carry out a systematic ablation study, with the detailed experimental configurations and results summarized in Table 3. In comparison to the baseline model, the proposed pseudo-label-guided multi-stage CL strategy boosts the average prediction accuracy by 7.4 PPs. The incorporation of the consistency retained module improves the average prediction accuracy by 2.2 PPs. The experimental results demonstrate that the model performance exhibits a consistent upward trend with the integration of an increasing number of modules. For instance, on the Caltech-4V dataset, the baseline model yields an ACC of 73.1% and an NMI of 67.5%. In contrast, the full-module integrated model exhibits substantial performance gains, attaining an ACC of 87.5% and an NMI of 78.3%. Consistent performance trends are observed across other benchmark datasets, with particularly prominent improvements on the NUS-Wide and Flickr datasets when all modules are incorporated. Specifically, the full-module model achieves an ACC of 53.9% and an NMI of 36.4% on the NUS-Wide dataset, and an ACC of 59.2% and an NMI of 38.2% on the Flickr dataset. The experimental results demonstrate that the model achieves optimal performance when all of its designed components are fully deployed.

3.6.2 Importance of pseudo-label guiding

To delve deeper into the significance of the pseudo-label-guided mechanism, we have conducted additional experiments on various multi-modal datasets. These experiments are performed both with and without the mechanism, and the outcomes are presented in Table 4. The results demonstrate that pseudo-label guiding significantly enhances the performance across most datasets. For instance, on the Caltech-3V dataset, the introduction of pseudo-label guiding leads to an improvement in ACC from 72.7% to 80.6%, and an enhancement in NMI from 63.4% to 69.9%. Similarly, on the Caltech-4V dataset, pseudo-label guiding results in a modest performance increase, with ACC rising from 87.0% to 87.5%, and NMI increasing from 76.7% to 78.3%. This may be attributed to

Table 2 Clustering results (%) in terms of ACC and NMI on multi-modal datasets

Method	Caltech-3V		Caltech-4V		NUS-Wide		Flickr		IAPR	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KM	46.3	31.3	54.6	46.7	26.8	16.7	40.9	22.5	38.9	17.2
Ncuts (TPAMI'00)	42.6	25.4	67.8	47.6	31.6	14.1	48.4	26.1	41.9	18.9
AmKM	46.9	31.5	44.9	30.6	26.8	15.2	41.0	21.6	40.4	17.0
AmNcuts (TPAMI'00)	43.7	25.5	41.8	24.9	30.4	16.1	48.2	26.2	42.2	18.9
CoregMVSC (NeurIPS'11)	54.4	45.3	64.9	54.5	26.3	16.8	41.0	26.8	35.1	18.4
RMKMC (IJCAI'13)	59.5	49.4	65.5	60.3	30.5	14.5	42.3	23.4	36.4	15.9
SwMC (IJCAI'17)	30.2	23.1	43.7	44.2	12.5	15.0	34.3	34.5	30.2	23.1
ONMSC (AAAI'20)	58.2	56.8	62.3	66.1	16.9	15.3	30.6	16.4	21.6	11.1
SMCMB (TBD'24)	67.2	54.5	74.4	67.0	32.4	21.8	52.8	34.1	34.8	16.4
EAMC (CVPR'20)	38.9	21.4	29.6	16.5	24.6	9.7	30.5	9.1	37.1	16.4
DEMC (InfoSci'21)	53.4	41.4	49.3	45.0	21.3	11.1	44.8	25.2	30.1	13.8
SiMVC (CVPR'21)	56.9	50.4	61.9	53.6	25.7	10.2	45.6	26.3	42.7	18.5
CoMVC (CVPR'21)	54.1	50.4	56.8	56.8	43.9	31.2	49.3	30.6	46.7	21.5
MFLVC (CVPR'22)	63.1	56.6	73.3	65.2	45.1	33.1	53.8	32.8	47.3	22.6
SEM (NeurIPS'23)	69.2	59.2	82.6	<u>75.3</u>	39.6	25.2	53.1	30.9	42.2	18.9
DIVIDE (AAAI'24)	60.9	53.8	64.3	57.9	36.3	26.0	52.3	33.5	45.6	23.0
SCMVC (TMM'24)	<u>75.9</u>	<u>66.3</u>	<u>84.4</u>	72.9	<u>48.8</u>	<u>34.2</u>	54.2	32.3	46.5	24.1
SSLNMVC (TMM'25)	64.4	58.3	82.1	72.8	38.9	28.3	51.2	33.0	46.4	24.0
MSDIB (AAAI'25)	74.6	66.2	66.4	55.2	47.6	33.4	<u>55.5</u>	<u>35.5</u>	<u>50.7</u>	<u>25.7</u>
Ours (MCMC-CR)	**80.6**	**69.9**	**87.5**	**78.3**	**53.9**	**36.4**	**59.2**	**38.2**	**54.6**	**30.4**

The bold and underlined values are the best and second best results, respectively

Table 3 Ablation study on different multi-modal datasets (%)

Method	Caltech-3V		Caltech-4V		NUS-Wide		Flickr		IAPR	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
$\mathcal{L}_{\text{DDC}}$	71.6	63.7	73.1	67.5	41.2	24.7	53.6	35.8	46.4	24.5
$\mathcal{L}_{\text{DDC}} + \mathcal{L}_{\text{MCL}}$	<u>76.4</u>	<u>69.5</u>	<u>86.5</u>	<u>76.8</u>	<u>50.2</u>	<u>35.8</u>	<u>57.6</u>	36.8	<u>52.4</u>	<u>29.8</u>
$\mathcal{L}_{\text{DDC}} + \mathcal{L}_{\text{Consistency}}$	75.5	67.0	78.7	73.1	41.4	30.0	54.6	<u>37.1</u>	46.8	26.9
All modules (proposed method)	**80.6**	**69.9**	**87.5**	**78.3**	**53.9**	**36.4**	**59.2**	**38.2**	**54.6**	**30.4**

The bold and underlined values are the best and second best results, respectively

Table 4 Experiments without/with pseudo-label guiding

Dataset	Pseudo-label guiding	ACC (%)	NMI (%)
Caltech-3V	–	72.7	63.4
	✓	**80.6**	**69.9**
Caltech-4V	–	87.0	76.7
	✓	**87.5**	**78.3**
NUS-Wide	–	53.6	35.5
	✓	**53.9**	**36.4**
Flickr	–	58.3	37.1
	✓	**59.2**	**38.2**
IAPR	–	54.1	29.8
	✓	**54.6**	**30.4**

– and ✓ denote the method without and with pseudo-label guiding, respectively. The better values are in bold

the fact that CL with pseudo-label guiding can effectively prevent intra-cluster samples from diverging, thereby enabling the model to learn more discriminative features.

3.6.3 Sensitivity analysis of initial K-means clustering

To evaluate the effectiveness and robustness of using K-means clustering in the original feature space to construct the negative sample mask matrix, we conduct a systematic sensitivity analysis experiment. The design of our comparative experiments to evaluate the impact of different $D_{i,j}$ construction strategies on the overall clustering performance of the proposed method is elaborated as follows.

1. Random $D_{i,j}$: The elements of the mask matrix $D_{i,j}$ are randomly assigned 0 or 1 with an equal probability, where the mask matrix $\boldsymbol{D}$ exerts a disruptive effect on the process of false negative pair elimination.

2. No $D_{i,j}$: abandon the mask matrix $\boldsymbol{D}$, where the elimination of false negative pairs is solely based on the clustering results derived from the DDC module.

3. K-means-based $D_{i,j}$: The mask matrix $\boldsymbol{D}$ is constructed by K-means clustering in the original feature space, which is also the strategy proposed in this paper.

4. Ground-truth $D_{i,j}$: The elements $D_{i,j}$ are set according to the real category labels of the samples. $D_{i,j}$=1 if and only if sample i and sample j belong to the same real category in all M modalities' original feature spaces, which represents the optimal theoretical state of false negative pair filtering.

From a theoretical perspective, the expected clustering performance of these four strategies follows a clear order: Ground-truth $D_{i,j}$>K-means-based $D_{i,j}$>No $D_{i,j}$>Random $D_{i,j}$. Ground-truth $D_{i,j}$ provides fully accurate prior knowledge and thus achieves the optimal performance. K-means-based $D_{i,j}$ offers reliable coarse-grained guidance, which is close to the ideal condition. No $D_{i,j}$ lacks the constraint from the original feature space and performs worse than the K-means-based strategy. Random $D_{i,j}$ introduces uncontrolled noise and severely disturbs false negative pair removal, leading to the worst performance. As expected, our experimental results are completely consistent with this theoretical ranking.

All the above strategies are implemented on the five benchmark datasets with identical experimental settings and hyperparameter configurations as the original method. This ensures the fairness and comparability of the obtained experimental results, which are summarized in Table 5. The experimental results demonstrate that the K-means-based $D_{i,j}$ strategy achieves clustering performance that is second only to the ground-truth $D_{i,j}$ on all benchmark datasets. Specifically, the average ACC of the K-means-based $D_{i,j}$ strategy across the five datasets is 0.48 PPs lower than that of the ground-truth $D_{i,j}$, and the average NMI is 0.40 PPs lower. In contrast, the Random $D_{i,j}$ strategy results in an average ACC reduction of 2.90 PPs and an average NMI reduction of 3.22 PPs relative to the ground-truth $D_{i,j}$ over the five datasets, which confirms that the rational construction of the mask matrix is the key to effective false negative pair elimination. The No $D_{i,j}$ strategy shows inferior performance compared with the K-means-based method, which indicates that the clustering results from the DDC module alone are insufficient to accurately identify false negative pairs at the early stage of training, and K-means clustering in the original feature space can provide reliable prior information for the negative sample selection module.

3.7 Parameter sensitivity

In the proposed MCMC-CR method, α and β are introduced to balance the contributions of the multi-stage CL loss ($\mathcal{L}_{\text{MCL}}$) and the consistency learning loss ($\mathcal{L}_{\text{Consistency}}$). To comprehensively evaluate the sensitivity of the model to these parameters, we perform extensive experiments on all datasets with diverse parameter configurations. Specifically, a grid search strategy is adopted to explore the optimal value combinations of α and β, where both parameters are sampled from the candidate set $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$. The ACC results under different parameter settings are presented in Fig. 2. As illustrated in the figure, the MCMC-CR method demonstrates robust performance across all datasets when $\beta \leq 10^{0}$, with only marginal performance fluctuations under most parameter configurations. However, a significant degradation in clustering ACC is observed when the value of β exceeds 10^{0}. This indicates that the MCMC-CR method is

Table 5 Clustering performance of different $D_{i,j}$ construction strategies on multi-modal datasets (%)

Method	Caltech-3V		Caltech-4V		NUS-Wide		Flickr		IAPR	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Random $D_{i,j}$	72.9	63.6	86.8	76.7	52.7	34.5	57.9	36.4	53.4	27.9
No $D_{i,j}$	76.8	67.4	87.4	77.9	53.4	35.9	59.1	37.8	54.2	28.3
K-means-based $D_{i,j}$	<u>80.6</u>	<u>69.9</u>	<u>87.5</u>	<u>78.3</u>	<u>53.9</u>	<u>36.4</u>	<u>59.2</u>	<u>38.2</u>	<u>54.6</u>	<u>30.4</u>
Ground-truth $D_{i,j}$	**81.2**	**70.3**	**88.4**	**78.9**	**54.2**	**36.7**	**59.6**	**38.4**	**54.8**	**30.9**

The bold and underlined values are the best and second best results, respectively

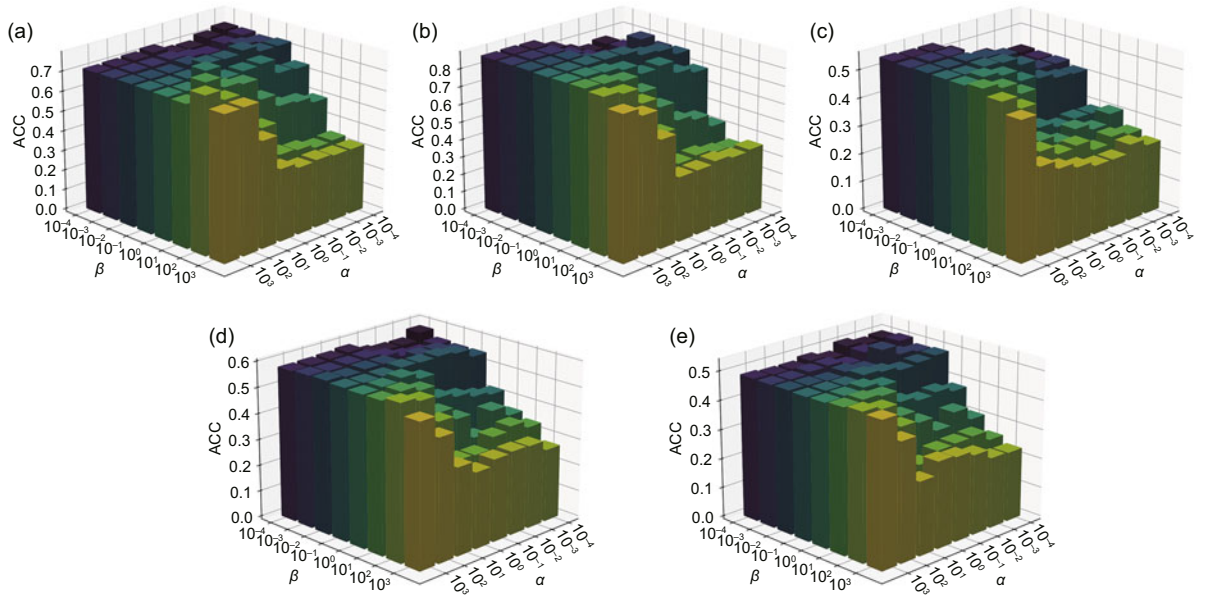

Fig. 2 Parameter analysis of α and β for MCMC-CR on different datasets: (a) Caltech-3V; (b) Caltech-4V; (c) NUS-Wide; (d) Flickr; (e) IAPR

relatively insensitive to variations in α and β within the range $\beta \leq 10^0$ but becomes highly sensitive to β when $\beta > 10^0$, which highlights a critical insight: The weight of consistency learning (governed by β) should not be set excessively high, as an overemphasis on consistency learning would lead to severe performance degradation.

To further evaluate the hyperparameter sensitivity of the proposed method, we perform grid search experiments on the candidate set $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ for the temperature hyperparameters τ_1 and τ_2, with corresponding results reported in Fig. 3. Specifically, τ_1 is employed in early/middle-stage CL and sample-level late-stage CL, while τ_2 is used for cluster-level late-stage CL. The goal of these experiments is to validate the model's hyperparameter sensitivity and determine the optimal hyperparameter values. Experimental results illustrate that the MCMC-CR method exhibits strong robustness against the variation of τ_1 and τ_2. In particular, the model maintains stable and superior performance when $\tau_1 \in [0.5, 1.0]$ and $\tau_2 \in [0.3, 1.0]$. This observation validates that our approach requires no tedious fine-grained adjustment for these two temperature parameters to guarantee effective CL in diverse multi-modal scenarios, thereby improving the practicality and generalization of the proposed method.

3.8 Visualization clustering results

To substantiate the efficacy of our proposed framework, we use the widely adopted t-distributed stochastic neighbor embedding (t-SNE) tool to visualize the clustering results of our method, MCMC-CR, on the Caltech-4V dataset. Fig. 4

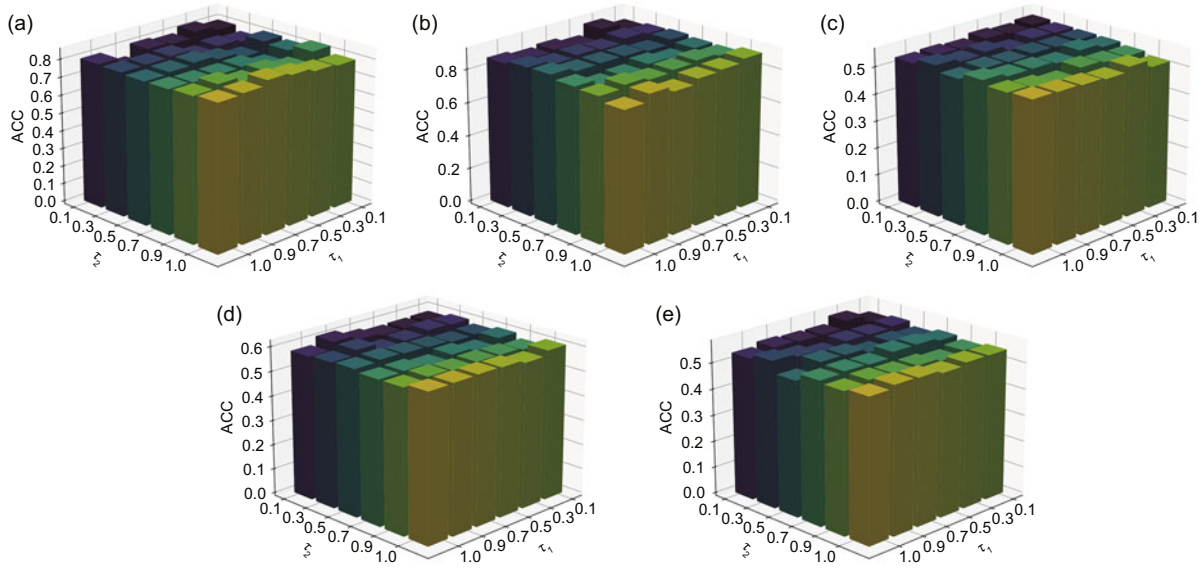

Fig. 3 Parameter analysis of τ_1 and τ_2 for MCMC-CR on different datasets: (a) Caltech-3V; (b) Caltech-4V; (c) NUS-Wide; (d) Flickr; (e) IAPR

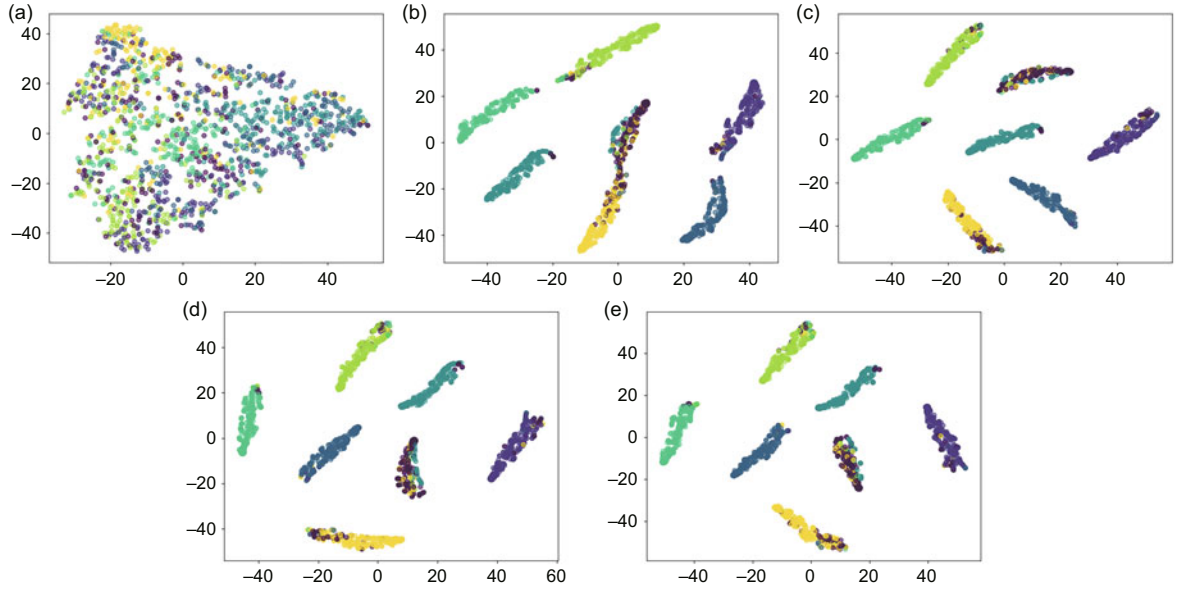

Fig. 4 Evolution of cluster assignments as the training goes on the Caltech-4V dataset: (a) epoch 0; (b) epoch 25; (c) epoch 50; (d) epoch 75; (e) epoch 100

illustrates the evolution of cluster assignments throughout training. From epoch 0 to epoch 100, the data points progress from an initial intermixed state to well-separated clusters. This demonstrates that MCMC-CR effectively captures the underlying data structure. As training advances, both intra-cluster compactness and inter-cluster separation are improved, leading to superior clustering performance.

3.9 Convergence analysis

To verify the convergence of the proposed method, we plot the convergence curves for the overall loss function, ACC, and NMI metrics across multiple dataset configurations in Fig. 5. These include the Caltech-3V, Caltech-4V, NUS-Wide, Flickr, and IAPR datasets. The results demonstrate that the loss value decreases rapidly within the first 20 training epochs before stabilizing. The ACC and NMI metrics show continuous improvement during the initial training phase, after which they also stabilize. This behavior clearly indicates that the proposed method exhibits strong convergence properties.

3.10 Runtime analysis

In this subsection, we present the runtime comparison between the compared deep MMC methods and the proposed method on five multi-modal datasets (Fig. 6). As the proposed method adopts a deep learning-based paradigm, we do not conduct comparisons with single/all-modal clustering algorithms and classic MMC approaches for fair comparison. It can be observed that the proposed method requires a moderately longer runtime compared with most deep MMC methods. This is attributed to the fact that the proposed method incorporates pseudo-label-guided multi-stage CL and demands additional computation time for consistency learning. Nevertheless, this computational overhead is worthwhile, as our method achieves satisfactory data clustering performance on the tested multi-modal datasets.

4 Conclusions

This paper introduces a novel multi-stage contrastive multi-modal clustering with consistency retained framework (MCMC-CR). It collaboratively preserves cross-modal consistency through early–middle–late CL across three levels: single-modal features, fused features, and cluster assignments. It enhances the accuracy of positive/negative sample pairs and mitigates the misleading effect of noisy pairs by jointly refining pseudo-labels based on the clustering results derived from both the original feature space and the fused feature space. The proposed method is designed for complete multi-modal data; however, it performs inadequately in complex scenarios with missing modalities or unaligned data. We plan to address these limitations in the future.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62576320), the Henan Provincial Outstanding Youth Science Fund Program (No. 252300421223), and the China Postdoctoral Science Foundation (Nos. 2024T170843 and 2023M743186).

Author contributions

Yanzheng WANG and Yujun WANG conceived the core idea, designed the overall framework of the study, led the implementation of experiments, drafted the paper, and revised it critically. Xin YANG and Fengshuo DAI participated in data preprocessing and experimental validation, providing valuable insights into data analysis. Xiaoheng JIANG contributed to the theoretical discussion and optimization of the algorithm design. Pei LV and Mingliang XU offered guidance on the technical feasibility and academic rigor of the research. Shizhe HU and Mingliang XU, as supervisors, proposed the research direction, supervised the entire project, and refined the paper with in-depth academic advice. All the authors reviewed and approved the final version of the paper.

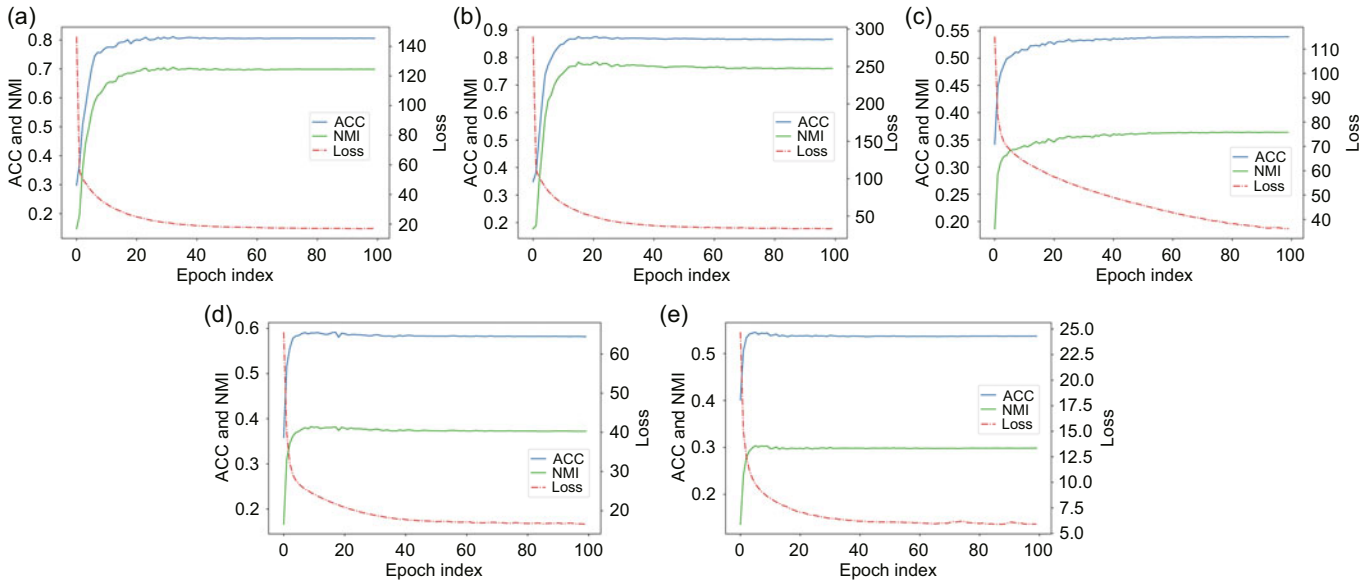

Fig. 5 Convergence analysis of MCMC-CR across Caltech-3V (a), Caltech-4V (b), NUS-Wide (c), Flickr (d), and IAPR (e) datasets

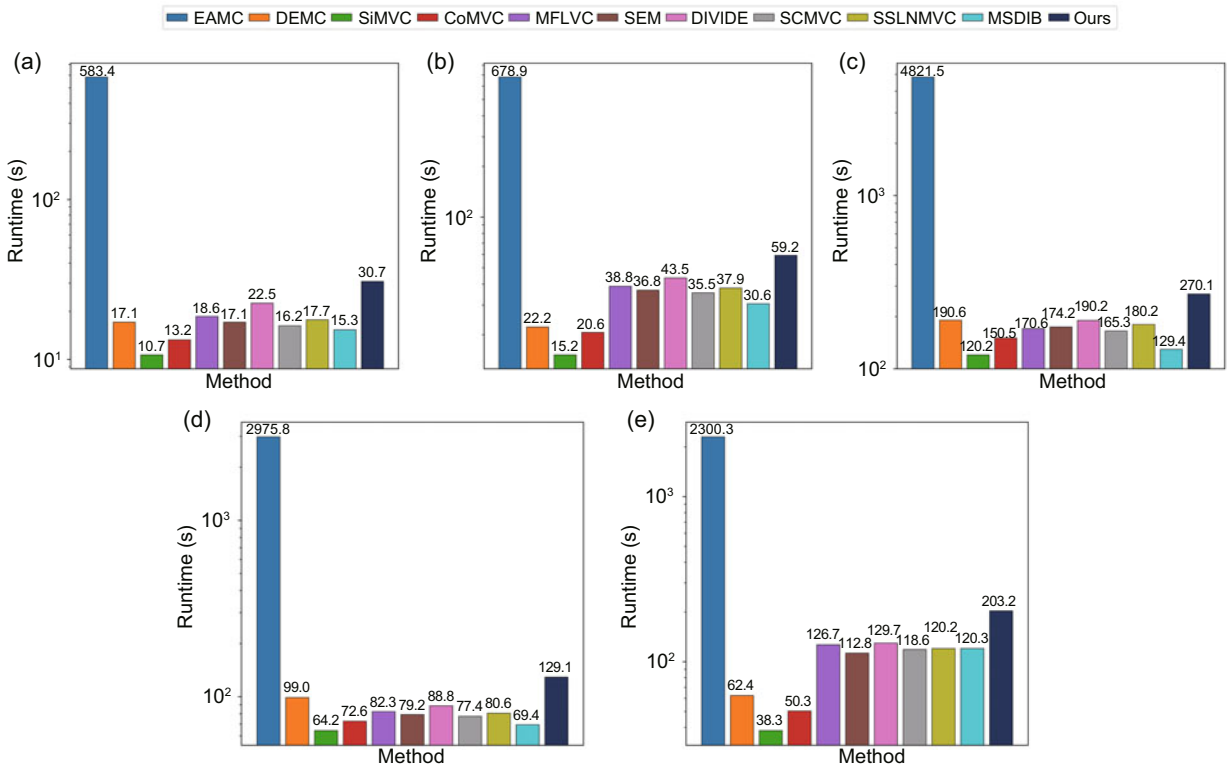

Fig. 6 Runtime comparison of the compared MMC methods and our methods on five multi-modal datasets: (a) Caltech-3V; (b) Caltech-4V; (c) NUS-wide; (d) Flickr; (e) IAPR

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Declaration on the use of generative AI tools

During the preparation of this work, the authors did not use any generative AI tools. All content was independently created, written, and revised by the authors, who take full responsibility for the integrity and accuracy of the paper.

References

Cai X, Nie F, Huang H, 2013. Multi-view K-means clustering on big data. Proc 23rd Int Joint Conf on Artificial Intelligence, p.2598-2604.

Chua TS, Tang J, Hong R, et al., 2009. NUS-WIDE: a real-world web image database from National University of Singapore. Proc ACM Int Conf on Image Video Retrieval, Article 48. https://doi.org/10.1145/1646396.1646452

Cui J, Li Y, Huang H, et al., 2024. Dual contrast-driven deep multi-view clustering. *IEEE Trans Image Process*, 33:4753-4764. https://doi.org/10.1109/TIP.2024.3444269

Dai Z, Wang G, Yuan W, et al., 2022. Cluster contrast for unsupervised person re-identification. 16th Asian Conf on Computer Vision, p.319-337. https://doi.org/doi/10.1007/978-3-031-26351-4-20

Friedman A, 1979. Framing pictures: the role of knowledge in automatized encoding and memory for gist. *J Exp Psychol-Gen*, 108(3):316-355. https://doi.org/10.1037/0096-3445.108.3.316

Grubinger M, Clough P, Müller H, et al., 2006. The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. https://www-i6.informatik.rwth-aachen.de/publications/download/34/Grubinger-LREC-2006.pdf [Accessed on Dec. 22, 2025]

Gu WJ, Zhu CM, 2024. Global and local combined contrastive learning for multi-view clustering. *Multim Syst*, 30(5):307. https://doi.org/10.1007/s00530-024-01512-8

Hartigan JA, Wong MA, 1979. Algorithm AS 136: a K-means clustering algorithm. *J R Stat Soc*, 28(1):100-108. https://doi.org/10.2307/2346830

Herath A, Kobti Z, 2024. Subtype-MMCC: multimodal contrastive clustering approach for cancer subtype discovery with multi-omics data. *Proc Comput Sci*, 246:696-705. https://doi.org/10.1016/j.procs.2024.09.488

Hu SZ, Zou GL, Zhang CY, et al., 2023. Joint contrastive triple-learning for deep multi-view clustering. *Inform Process Manag*, 60(3):103284. https://doi.org/10.1016/j.ipm.2023.103284

Hu SZ, Zhang CK, Zou GL, et al., 2025a. Deep multiview clustering by pseudo-label guided contrastive learning and dual correlation learning. *IEEE Trans Neur Netw Learn Syst*, 36(2):3646-3658. https://doi.org/10.1109/TNNLS.2024.3354731

Hu SZ, Fan JH, Zou GL, et al., 2025b. Multi-aspect self-guided deep information bottleneck for multi-modal clustering. Proc 39th AAAI Conf on Artificial Intelligence, p.17314-17322. https://doi.org/10.1609/aaai.v39i16.33903

Huiskes MJ, Lew MS, 2008. The MIR Flickr retrieval evaluation. Proc 1st ACM Int Conf on Multimedia Information Retrieval, p.39-43. https://doi.org/10.1145/1460096.1460104

Joyce JM, 2011. Kullback–Leibler Divergence. Springer Berlin, Heidelberg, p.720-722. https://doi.org/10.1007/978-3-642-04898-2_327

Kampffmeyer M, Løkse S, Bianchi FM, et al., 2019. Deep divergence-based approach to clustering. *Neur Netw*, 113:91-101. https://doi.org/https://doi.org/10.1016/j.neunet.2019.01.015

Ke GZ, Hong ZY, Zeng ZQ, et al., 2021. CONAN: contrastive fusion networks for multi-view clustering. Proc IEEE Int Conf on Big Data, p.653-660. https://doi.org/10.1109/BigData52589.2021.9671851

Kingma DP, Ba J, 2017. Adam: a method for stochastic optimization. https://doi.org/10.48550/arXiv.1412.6980

Kumar A, Rai P, Daumé H III, 2011. Co-regularized multi-view spectral clustering. Proc 25th Int Conf on Neural Information Processing Systems, p.1413-1421.

Lao JH, Huang D, Wang CD, et al., 2024. Towards scalable multi-view clustering via joint learning of many bipartite graphs. *IEEE Trans Big Data*, 10(1):77-91. https://doi.org/10.1109/TBDATA.2023.3325045

Li FF, Fergus R, Perona P, 2004. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. Proc IEEE Conf on Computer Vision and Pattern Recognition Workshop, p.178. https://doi.org/10.1109/CVPR.2004.383

Lin K, Wang D, Xia FZ, et al., 2018. Device clustering algorithm based on multimodal data correlation in cognitive Internet of Things. *IEEE Int Things J*, 5(4):2263-2271. https://doi.org/10.1109/JIOT.2017.2728705

Lou ZZ, Xue H, Wang YZ, et al., 2025. Parameter-free deep multi-modal clustering with reliable contrastive learning. *IEEE Trans Image Process*, 34:2628-2640. https://doi.org/10.1109/TIP.2025.3562083

Lu YD, Lin YJ, Yang MX, et al., 2024. Decoupled contrastive multi-view clustering with high-order random walks. Proc 38th AAAI Conf on Artificial Intelligence, p.14193-14201. https://doi.org/10.1609/aaai.v38i13.29330

Nie FP, Li J, Li XL, 2017. Self-weighted multiview clustering with multiple graphs. Proc 26th Int Joint Conf on Artificial Intelligence, p.2564-2570.

Pan L, Sun GD, Chang BF, et al., 2023. Visual interactive image clustering: a target-independent approach for configuration optimization in machine vision measurement. *Front Inform Technol Electron Eng*, 24(3):355-372. https://doi.org/10.1631/FITEE.2200547

Shen DG, Ip HHS, 1999. Discriminative wavelet shape descriptors for recognition of 2-D patterns. *Patt Recog*, 32(2):151-165. https://doi.org/10.1016/S0031-3203(98)00137-X

Shi JB, Malik J, 2000. Normalized cuts and image segmentation. *IEEE Trans Patt Anal Mach Intell*, 22(8):888-905. https://doi.org/10.1109/34.868688

Tian YL, Krishnan D, Isola P, 2020. Contrastive multiview coding. Proc 16th European Conf on Computer Vision, p.776-794. https://doi.org/10.1007/978-3-030-58621-8-45

Tong Y, 2025. Research on deep clustering and multimodal deep learning service recommendation system for large-scale user data. 3rd Int Conf on Integrated Circuits and Communication Systems, p.1-5. https://doi.org/10.1109/ICICACS65178.2025.10967931

Trosten DJ, Løkse S, Jenssen R, et al., 2021. Reconsidering representation alignment for multi-view clustering. Proc IEEE/CVF Conf on Computer Vision and Pattern, p.1255-1265. https://doi.org/10.1109/CVPR46437.2021.00131

Wolf L, Hassner T, Taigman Y, 2008. Descriptor based methods in the wild. Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition.

Wu JX, Rehg JM, 2011. CENTRIST: a visual descriptor for scene categorization. *IEEE Trans Patt Anal Mach Intell*, 33(8):1489-1501. https://doi.org/10.1109/TPAMI.2010.224

Wu S, Zheng Y, Ren YZ, et al., 2024. Self-weighted contrastive fusion for deep multi-view clustering. *IEEE Trans Multim*, 26:9150-9162. https://doi.org/10.1109/TMM.2024.3387298

Xu J, Ren YZ, Li GF, et al., 2021. Deep embedded multi-view clustering with collaborative training. *Inform Sci*, 573:279-290. https://doi.org/10.1016/j.ins.2020.12.073

Xu J, Tang HY, Ren YZ, et al., 2022. Multi-level feature learning for contrastive multi-view clustering. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.16030-16039. https://doi.org/10.1109/CVPR52688.2022.01558

Xu J, Chen S, Ren YZ, et al., 2023. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. Proc 37th Int Conf on Neural Information Processing Systems, Article 54.

Yan WQ, Yang TY, Tang C, 2025. Self-supervised semantic soft label learning network for deep multi-view clustering. *IEEE Trans Multim*, 27:4971-4983. https://doi.org/10.1109/TMM.2025.3543075

Yang KL, Zhang TL, Alhuzali H, et al., 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Trans Affect Comput*, 14(4):3269-3280. https://doi.org/10.1109/TAFFC.2023.3243463

Yin YJ, Ruan H, Chen Y, et al., 2025. Prototypical clustered federated learning for heart rate prediction. *Front Inform Technol Electron Eng*, 26(10):1896-1912. https://doi.org/10.1631/FITEE.2500062

Zhang FD, Kuang K, Chen L, et al., 2023. Federated unsupervised representation learning. *Front Inform Technol Electron Eng*, 24(8):1181-1193. https://doi.org/10.1631/FITEE.2200268

Zhou RW, Shen YD, 2020. End-to-end adversarial-attention network for multi-modal clustering. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.14607-14616. https://doi.org/10.1109/CVPR42600.2020.01463

Zhou SH, Liu XW, Liu JY, et al., 2020. Multi-view spectral clustering with optimal neighborhood Laplacian matrix. Proc 34th AAAI Conf on Artificial Intelligence, p.6965-6972. https://doi.org/10.1609/aaai.v34i04.6180