SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

July 2026, Vol. 69, Iss. 7, 172104:1–172104:15
https://doi.org/10.1007/s11432-025-4667-2

Cross-level interaction and multi-granularity contrastive learning for multi-view clustering

Shanghui DENG[1], Chang TANG[2*], Xiao ZHENG[3], Yuanyuan LIU[1], Kun SUN[1] & Xinwang LIU[4]

[1]*School of Computer Science, China University of Geosciences (Wuhan), Wuhan 430074, China*
[2]*School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, China*
[3]*School of Computer Science, Hubei University of Technology, Wuhan 430068, China*
[4]*School of Computer, National University of Defense Technology, Changsha 410073, China*

Received 13 May 2025/Revised 21 August 2025/Accepted 31 October 2025/Published online 16 June 2026

Abstract Contrastive learning (CL) is extensively applied in multi-view clustering by pulling positive samples closer and pushing negative samples apart. However, most existing deep contrastive multi-view clustering (DCMVC) methods are constrained to a single perspective, either at the feature or/and semantic level, ignoring the rich and crucial information between these two levels at intermediate granularities. Additionally, these methods concentrate solely on contrastive within the same level, overlooking potential interactions across different levels. This limitation restricts the representation capability of DCMVC, subsequently impacting clustering performance. To this end, we propose a novel framework dubbed CLMGC: cross-level interaction and multi-granularity contrastive learning for multi-view clustering. Specifically, we input the latent representations of each encoder into feature-level and semantic-level CL. In semantic-level CL, we propose a novel subdivision of the semantic level into two branches: fine-grained and coarse-grained semantic clusters. To further enhance the hierarchical richness of information, we introduce fine-grained dual contrastive mechanisms, including cross-level and self-instance contrast mechanisms, which connect feature-level with fine-grained semantic CL. This design enhances information transfer between different levels and significantly improves the discriminative capability of the fine-grained semantic cluster, thus optimizing the overall performance of CLMGC. Experimental results from six multi-view datasets demonstrate the superiority of the CLMGC algorithm compared with other state-of-the-art methods. The demo code of this work is publicly available at https://shanghui-deng.github.io.

Keywords multi-view clustering, cross-level interaction, multi-granularity, dual contrastive mechanisms, contrastive multi-view clustering

Citation Deng S H, Tang C, Zheng X, et al. Cross-level interaction and multi-granularity contrastive learning for multi-view clustering. Sci China Inf Sci, 2026, 69(7): 172104, https://doi.org/10.1007/s11432-025-4667-2

1 Introduction

With the rapid development of multimedia applications, a large amount of data have been collected from various sources or described with different attributes. For example, video data contain images captured by various cameras, audio materials, and text narration. Unfortunately, this data often suffer from missing labels. To uncover consistency and complementary information across multiple views in an unsupervised environment, multi-view clustering (MVC) [1–8] technology emphasizes the integration of various data sources for a more comprehensive understanding of the underlying phenomena.

Exiting MVC methods can be roughly categorized into traditional methods and deep methods. Traditional MVC methods utilize classic machine learning techniques for clustering tasks and are further refined into multiple branches, such as subspace methods [9–13], matrix factorization methods [14–17], and graph methods [18–23]. Nevertheless, many traditional MVC methods are constrained by hand-designed feature extraction and linear embedding models, making it challenging to effectively reveal the nonlinear structures hidden in complex view data.

Recently, deep MVC technology has garnered significant attention and recognition due to the exceptional capabilities of deep neural networks (DNNs) in nonlinear feature learning [24–26]. Deep MVC methods can be broadly classified into two categories: two-stage methods [27–29] and one-stage methods [30–32]. The two-stage deep MVC methods employ the powerful feature ability of DNNs to thoroughly explore the underlying potential features, followed by applying traditional clustering algorithms such as K-Means [33] or spectral clustering [34] to achieve the final clustering. The one-stage deep MVC methods demonstrate higher integration and efficiency, seamlessly

* Corresponding author (email: tangchang@hust.edu.cn)

integrating feature learning and clustering tasks into a unified framework, thereby facilitating an end-to-end process from feature extraction to the output of clustering results.

Although deep MVC methods have made significant progress, most existing methods still focus on leveraging consistency to reveal the underlying feature consistency across views [28, 35–39]. In contrast, in pursuing feature space consistency, these methods overlook a critical issue: the reconstruction target may inadvertently result in the redundant reconstruction of crucial features, introducing unnecessary private information that interferes with the effectiveness of information extraction. To this end, the deep contrastive multi-view clustering (DCMVC) [40–43] method has emerged, offering breakthroughs in deep MVC through its unique perspective and methodology. The core of the DCMVC method involves deep mining and refining the alignment represented by each view, aiming to extract consensus information from multiple perspectives. Different views of the same sample are treated as positive, while non-homologous views are considered negative in DCMVC [44–46]. With this setup, the DCMVC method can effectively reduce the distance between positive samples and increase the distance from negative samples, thereby creating a more accurate and expressive latent representation. The contrastive learning strategy significantly enhances consistency between views and improves the accuracy of information extraction.

However, existing DCMVC methods, while effective, often operate under one of two dominant yet relatively independent paradigms, creating a significant gap in how information is leveraged across different levels of abstraction. The first paradigm is instance-level contrast. Its core objective is to learn view-invariant instance representations by pulling different views of the same instance together (as positive pairs) while pushing views of different instances apart (as negative pairs) in the feature space [47, 48]. While powerful for learning discriminative features, this approach is fundamentally "cluster-agnostic". It treats each instance as an isolated entity, and the representation learning process does not explicitly leverage the underlying group structure of the data, while clustering is often treated as a subsequent downstream task. The second paradigm is semantic-level contrast. These methods operate on a higher level of abstraction by directly contrasting cluster assignments or prototypes across views [45, 49]. The goal is to enforce global consistency in the final clustering results. While this ensures macro-level agreement, the connection between the low-level instance features and these high-level semantic assignments can be indirect, potentially failing to fully utilize fine-grained instance-level details to guide the representation learning process. This dichotomy constitutes the overlooked potential interactions across different levels that we aim to address. The crucial missing piece is a mechanism that establishes a direct, structured dialogue between the instance-feature space and the semantic-group space. In prior work, feature learning is not explicitly guided by a rich semantic structure, and conversely, the clustering process does not fully benefit from a feature space that has been "custom-built" for the grouping task.

To bridge this gap and establish this crucial interaction, we propose a novel framework dubbed CLMGC (as shown in Figure 1). The central innovation of CLMGC is the construction of a "semantic bridge" to facilitate a symbiotic relationship between instance-level features and cluster-level semantics. Within this framework, feature learning is no longer blind but is explicitly guided by its semantic destination, while the clustering process is built upon a more discriminative and better-structured feature space. To achieve this, our framework pioneers a multi-granularity architecture. We subdivide the semantic level into two branches: a coarse-grained branch for ensuring global consistency in the final clustering task, and a fine-grained branch that serves as the core of our "semantic bridge". This fine-grained semantic cluster is then empowered by a novel dual contrastive mechanism: (1) a cross-level contrast that explicitly links the instance-level feature branch with this fine-grained semantic branch, creating the vital instance-to-group connection; and (2) a self-instance contrast that operates within the fine-grained branch to enhance its own structural integrity and discriminative power. By synergistically integrating feature-level, coarse-grained, and fine-grained dual contrastive learning, CLMGC trains its network in a self-supervised manner to achieve highly effective clustering.

Compared with previous work, our main contributions to this paper are listed as follows.

- We propose a cross-level semantic interaction method for multi-granularity contrastive multi-view clustering, termed CLMGC, which aims to uncover and exploit comprehensive semantic information between feature-level and semantic-level contrasts.
- We incorporate a fine-grained semantic cluster branch as a bridge between feature-level and semantic-level contrastive learning, intending to extract and integrate multi-level semantic information to enhance cross-level semantic interaction and understanding.
- We develop the fine-grained dual contrastive mechanism: the cross-level and the self-instance contrast mechanism. The former enhances deep semantic interaction between feature and semantic levels, improving the hierarchy and richness of information interaction. At the same time, the latter focuses on strengthening the discriminative ability of the fine-grained semantic cluster.

The rest of this paper is organized as follows. We review existing multi-view clustering methods in Section 2.

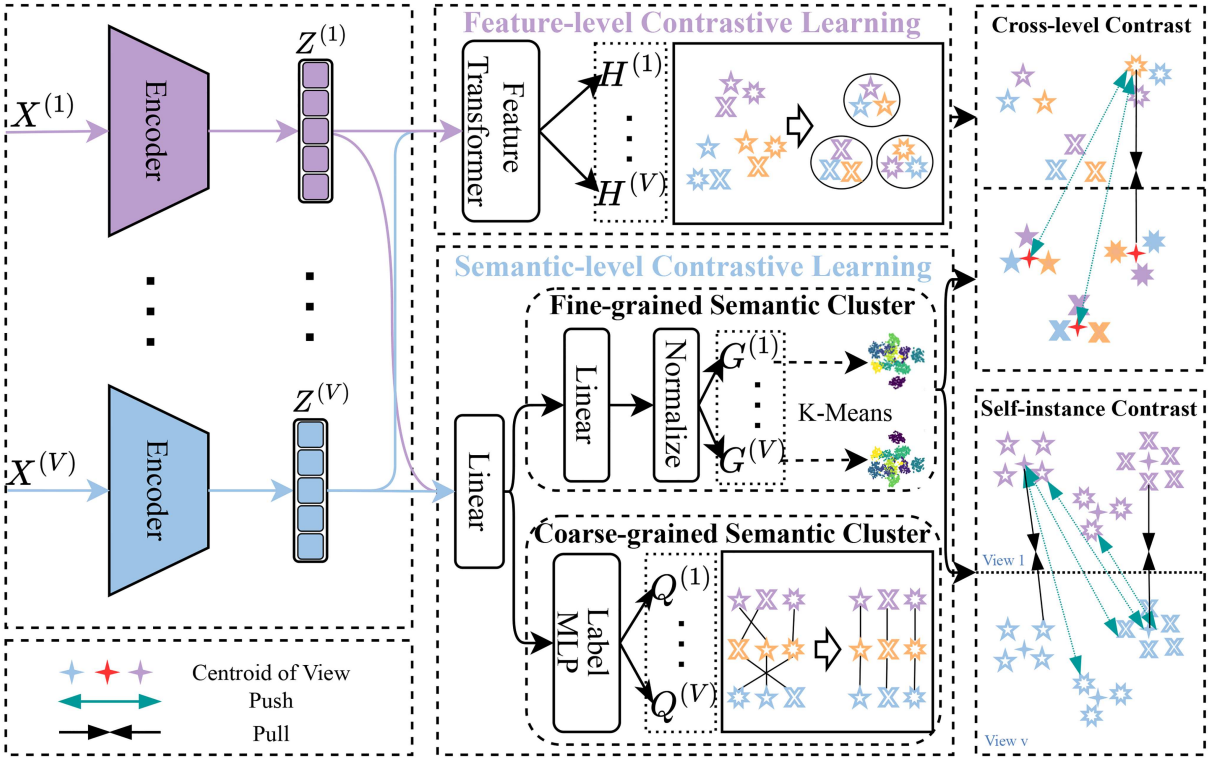

Figure 1 (Color online) Framework overview of CLMGC. CLMGC processes multi-view data $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$ into low-dimensional $\{\mathbf{Z}^{(v)}\}_{v=1}^{V}$, then extracts high-level features $\{\mathbf{H}^{(v)}\}_{v=1}^{V}$ and cluster assignments $\{\mathbf{Q}^{(v)}\}_{v=1}^{V}$ via a Transformer and MLP for feature and semantic contrastive learning. The fine-grained dual contrastive mechanism, with cross-level and self-instance components, further refines these features by promoting inter-level interaction and enhancing fine-grained semantic cluster distinction, respectively. This design enables effective multi-view data structure utilization.

Section 3 introduces a detailed introduction to our proposed model. Extensive experiments and comparisons to the state-of-the-art techniques are reported in Section 4. Section 5 summarizes the entire paper.

2 Related work

2.1 Multi-view clustering

Existing MVC methods can be roughly classified into four categories. (1) Subspace-based MVC [9–11]. The core of this type of method is to explore and learn the shared subspace representation across multiple views. Ref. [50] leveraged the complementary characteristics of views to achieve a more accurate and robust latent subspace representation. Liu et al. [51] introduced a novel integration of anchor learning and graph construction to develop a comprehensive framework. Notably, the algorithm directly derives clustering results from graph connectivity constraints, achieving efficient and accurate clustering analysis. (2) Matrix factorization-based MVC [14–16, 52–54] that aligns with the relaxed form of the K-means algorithm. Non-negative matrix factorization is a notable representative of this category. By decomposing each view into a low-rank matrix, this method effectively aggregates data in a low-dimensional space. Wei et al. [55] proposed a deep matrix factorization approach that decomposes the multi-view data matrix into multiple representation subspaces layer by layer, which enables a deep exploration of the data's underlying structure. (3) Graph-based MVC [18–20, 23, 56] focuses on accurately constructing graph structures to capture and preserve sample adjacency relationships. Ref. [57] utilized graph autoencoders to learn latent cluster representations and used an information-rich graph view as input to reconstruct latent representations into multiple graph views, accurately capturing complex data structures. More specifically, Ref. [52] advanced multi-view spectral clustering by leveraging tailored tensor low-rank representation for higher-order correlation capture. (4) Deep MVC [27–29, 58–60]. With the rapid advancement of deep learning, deep MVC methods are gaining increased attention from researchers. Deep MVC can be further categorized into two-stage and single-stage methods based on processing flow [30]. These methods leveraged the strengths of DNNs in representation learning to explore latent clustering patterns in multi-view data, offering new perspectives and powerful tools for cluster

analysis. Beyond these specific multi-view paradigms, general strategies like ensemble clustering [61] have also advanced the field by enhancing the generalization and robustness of clustering algorithms.

2.2 Contrastive multi-view clustering

Contrastive learning (CL) [47,62,63], as a cutting-edge unsupervised representation learning method, is fundamentally based on maximizing the similarity between positive samples and minimizing the relation between negative samples in the feature space. In computer vision, CL has gained widespread recognition for its exceptional ability in feature learning. Li et al. [49] pioneered an online image contrastive clustering method that applied CL at both the instance and cluster levels, showcasing its practical value and broad applicability. Similarly, CL has demonstrated significant potential in multi-view data processing. Ref. [30] adopted a fusion-free strategy to explore different levels of information in the original features, achieving multi-level feature learning through contrast. Chen et al. [45] proposed a cross-view contrastive learning method focused on learning view-invariant representations and generating clustering results by contrasting clustering assignments across views. In [64], researchers proposed a dual mutual information-constrained clustering method that minimizes mutual information across all dimensions while maximizing it between similar instance pairs, thereby improving clustering performance. Despite its success in various applications, most DCMVC research has primarily concentrated on feature or/and semantic-level contrastive, overlooking the rich information at intermediate granularities. For instance, instance-centric methods (e.g., DealMVC) learn cluster-agnostic features before grouping, while semantic-centric methods (e.g., Chen et al.) contrast cluster assignments, ignoring structural details. Both miss fine-grained semantic guidance for feature learning. This limitation restricts their ability to capture complex data structures effectively. To address this gap, we propose a CLMGC to enable more accurate and efficient clustering for multi-view data.

3 The proposed method

In this paper, we propose a CLMGC framework. In this section, we present an overview of the CLMGC model, followed by a detailed description of its main components, and conclude with the introduction of the objective function employed by the model.

3.1 Overview of CLMGC

In this section, we present a detailed explanation of our proposed CLMGC model. The framework overview of CLMGC is shown in Figure 1. The CLMGC model consists of four main components: representation extraction, feature-level contrastive learning, coarse-grained semantic-level contrastive learning, and fine-grained dual contrastive mechanism. To be specific, we first employ independent encoders to extract latent feature representations for each view, followed by constructing two parallel learning branches: a feature-level contrastive learning branch and a semantic-level contrastive learning branch. In the feature-level contrastive learning branch, feature transformers capture and enhance the latent feature representations to obtain higher-level latent representations. Feature-level contrastive learning is applied to these high-level latent representations to prevent model collapse and thoroughly explore common semantic information across views. As for the semantic-level contrastive learning branch, we further refine it into the coarse-grained semantic cluster and the fine-grained semantic cluster. Label assignments are derived from each latent representation at the coarse-grained semantic cluster, and cluster contrastive learning is performed to ensure consistent representation clustering. At the fine-grained cluster, latent representations are mapped to more compact fine-grained clustering spaces. We develop a fine-grained dual contrastive mechanism consisting of cross-level and self-instance contrast mechanisms. The cross-level contrast mechanism connects feature-level and semantic-level contrast learning to facilitate cross-level interaction. This mechanism fosters interaction between the two levels, enabling deep exploration of rich intermediate granularity information. Additionally, we introduced the self-instance contrast mechanism within the fine-grained semantic cluster. This innovative design enhances the discriminative power of the fine-grained semantic cluster, optimizing the model's overall clustering performance.

3.2 Representation extraction and enhancement

As a highly regarded unsupervised learning model, the encoder significantly enhances model performance by extracting key features from input data. Its multi-layer structure with nonlinear activation functions grants the encoder a strong capability to perform nonlinear transformations, enabling it to capture complex nonlinear relationships in the input data. Therefore, CLMGC leverages encoder technology to extract deep latent representations from multi-view data. Specifically, consider a multi-view dataset $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$ containing N samples across V views, where

$\mathbf{X}^{(v)} = \{\mathbf{X}_1^{(v)}; \mathbf{X}_2^{(v)}; \cdots; \mathbf{X}_N^{(v)}\} \in \mathbb{R}^{N \times D_v}$, D_v represents the dimension of the original features in the v-th view. For the v-th view, CLMGC utilizes independent encoder $f_{Encoder}^{(v)}$ to learn their respective d_v-dimensional feature representations $\mathbf{Z}^{(v)} \in \mathbb{R}^{N \times d_v}$:

$$\mathbf{Z}^{(v)} = f_{Encoder}^{(v)}(\mathbf{X}^{(v)}), \tag{1}$$

where d_v is the dimension of latent representation, $\mathbf{Z}^{(v)}$ represents the latent low-dimensional representations extracted from the feature attribute data of $\mathbf{X}^{(v)}$, and $f_{Encoder}^{(v)}$ denotes the encoder function.

We employ two distinct architectures, a Transformer and an MLP, to serve the complementary goals of our parallel branches. For the feature-level branch, a Transformer is used to capture rich, inter-sample relationships via its self-attention mechanism. This is essential for enhancing the feature representations $\mathbf{H}^{(v)}$ so they are globally aware and highly discriminative, a prerequisite for effective instance-level contrast. Conversely, for the semantic-level branches, the task is to perform a direct mapping from features to cluster assignments. For this, we use a shared MLP, which serves as a powerful non-linear function approximator capable of learning this mapping for each instance independently. This deliberate architectural choice allows each branch to leverage the most suitable tool for its specific function: relational enhancement versus direct semantic mapping.

To capture global structural relationships among the low-level features $\{\mathbf{Z}^{(v)}\}_{v=1}^{V}$, we employ a standard self-attention block, a core component of the Transformer architecture [65]. This allows the representation of each sample to be enhanced by aggregating information from highly correlated samples within the same view. The resulting high-level features $\mathbf{H}^{(v)}$ are computed as

$$\mathbf{H}^{(v)} = softmax\left(\frac{(\mathbf{Z}^{(v)}\mathbf{W}_Q)(\mathbf{Z}^{(v)}\mathbf{W}_K)^{\mathrm{T}}}{\sqrt{D_v}}\right)\mathbf{Z}^{(v)}\mathbf{W}_V, \tag{2}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable matrices.

3.3 Feature-level contrastive learning

The primary objective of our feature-level contrastive learning module is to perform instance-level discrimination. Positive pairs are formed by different views of the same data instance, while negative pairs consist of views from different instances. This strategy, operating on the high-level features $\mathbf{H}^{(v)}$, aims to learn an embedding space where individual samples are maximally distinguished from one another, thereby capturing fine-grained, instance-specific information.

Specifically, each high-level feature $\mathbf{h}_i^v \in \mathbf{H}^{(v)}$, there are $(VN-1)$ feature pairs, i.e., $\{\mathbf{h}_i^v, \mathbf{h}_j^n\}_{j=1,\ldots,N}^{n=1,\ldots,V}$. Among these feature pairs, $\{\mathbf{h}_i^v, \mathbf{h}_i^n\}_{n \neq m}$ constitutes $(V-1)$ positive feature pairs. The remaining $V(N-1)$ feature pairs are considered negative feature pairs. Following the NT-Xent loss function [47], we employ cosine distance to measure the similarity between two features:

$$s(\mathbf{h}_i^v, \mathbf{h}_j^n) = \frac{\langle \mathbf{h}_i^v, \mathbf{h}_j^n \rangle}{\|\mathbf{h}_i^v\| \, \|\mathbf{h}_j^n\|}, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes dot product operator. Afterward, the feature contrastive loss between high-level features $\mathbf{H}^{(v)}$ and $\mathbf{H}^{(n)}$ is formulated as

$$\ell_{fc}^{(vn)} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s(\mathbf{h}_i^v, \mathbf{h}_i^n)/\tau_F}}{\sum_{j=1}^{N}\sum_{m=v,n} e^{s(\mathbf{h}_i^v, \mathbf{h}_j^m)/\tau_F} - e^{1/\tau_F}}, \tag{4}$$

where τ_F is the temperature parameter. In this paper, we propose a cumulative multi-view feature contrast loss function that integrates relevant information across all views. The function is defined as follows:

$$\mathcal{L}_{FC} = \frac{1}{2}\sum_{v=1}^{V}\sum_{n \neq v} \ell_{fc}^{(vn)}. \tag{5}$$

Ensuring goal consistency on $\{\mathbf{H}^{(v)}\}_{v=1}^{V}$ allows for deeper exploration of common semantic information across views. This strategy brings the clusters of high-level features closer to the true semantic clusters. Intuitively, semantic information represents a high-level conceptual category that naturally excludes irrelevant noise. Thus, high-level features are closely aligned in the same semantic cluster, forming a compact and dense distribution. In addition to focusing on the finest-grained feature-level contrastive, we also explore the distinguishability at the feature group level to uncover the intrinsic connections between coarse-grained and fine-grained semantic clusters across different granularities.

3.4 Coarse-grained semantic-level contrastive learning

Many contemporary deep contrastive clustering methods, such as the pioneering Contrastive Clustering [49], successfully integrate instance-level and semantic-level objectives. However, a common feature of these approaches is their reliance on a single semantic granularity, where contrast is performed on the final K cluster assignments. While effective for enforcing global consistency, this single, coarse-grained level provides a relatively distant supervisory signal to the instance-level feature learning process. Our central motivation is to bridge this gap. We hypothesize that a richer, intermediate semantic structure can facilitate a more direct and effective interaction between the feature and semantic spaces.

To realize this vision, our semantic-level contrastive learning in the CLMGC framework is divided into two synergistic branches. The first branch, detailed in this section, focuses on the coarse-grained semantic cluster. This aligns with standard practice to ensure overall clustering stability and serves as the foundation for our framework. The second branch, the fine-grained semantic cluster, delves deeper into subtle cluster structures through over-clustering. It functions as a "semantic scaffolding" that provides crucial support for our novel dual contrastive mechanism. The following section details the implementation process for the coarse-grained clustering branch.

Formally, the low-level features $\{\mathbf{Z}^{(v)}\}_{v=1}^{V}$ extracted from the encoder are input into a shared label MLP to obtain the cluster assignments $\{\mathbf{Q}^{(v)} \in \mathbb{R}^{N\times K}\}_{v=1}^{V}$ for each view, represented by the process $f_{\vartheta}(\mathbf{Z}^{(v)})$, K is the number of cluster, and ϑ is the parameter of the label MLP. The last layer of the shared label MLP is equipped with a Softmax operation to ensure the output is a probability distribution. In detail, $q_{ij}^{(v)}$ represents the probability that the i-th sample is assigned to the j-th cluster in the v-th view.

In practical applications, misleading view-specific information may result in incorrect cluster label assignments for some views of a sample. To improve the model's robustness, it is essential to ensure consistent clustering, meaning the same cluster labels across all views accurately represent the same semantic group. In other word, $\{\mathbf{Q}_{\cdot j}^{(v)}\}_{v=1}^{V}$ require to remain consistent. To achieve this, similar to feature-level contrastive learning, we introduce contrastive learning in cluster assignments to reinforce semantic consistency. For the v-th view, there are $(VK-1)$ potential label pairs for the same cluster labels $\mathbf{Q}_{\cdot j}^{(v)}$, denoted as $\{\mathbf{Q}_{\cdot j}^{(v)}, \mathbf{Q}_{\cdot k}^{(n)}\}_{k=1,\ldots,K}^{v=1,\ldots,V}$. $\{\mathbf{Q}_{\cdot j}^{(v)}, \mathbf{Q}_{\cdot j}^{(n)}\}_{v\neq n}$ from $(M-1)$ positive label pairs and the remaining $V(K-1)$ pairs are considered negative label pairs. To quantify and optimize label consistency between $\mathbf{Q}^{(v)}$ and $\mathbf{Q}^{(n)}$, we define the semantic-level contrastive loss as follows:

$$\ell_{sc}^{(vn)} = -\frac{1}{K}\sum_{j=1}^{K}\log\frac{e^{s(\mathbf{Q}_{\cdot j}^{(v)},\mathbf{Q}_{\cdot j}^{(n)})/\tau_S}}{\sum_{k=1}^{K}\sum_{m=v,n}e^{s(\mathbf{Q}_{\cdot j}^{(v)},\mathbf{Q}_{\cdot k}^{(m)})/\tau_S} - e^{1/\tau_S}}, \tag{6}$$

where τ_S represents the temperature parameter. To avoid all instances being inappropriately classified into a single cluster, we introduce a regularization constraint defined as follows:

$$\ell_{sa}^{(v)} = \sum_{k=1}^{K} p_k^{(v)}\log p_k^{(v)}, \tag{7}$$

where $p_k^{(v)} = \frac{1}{N}\sum_{i=1}^{N} q_{ik}^{(v)}$. This term serves as a loss metric to evaluate cross-view consistency. As a result, the consistency objective of coarse-grained semantic contrastive learning can be explicitly defined as

$$\mathcal{L}_{SC} = \frac{1}{2}\sum_{v=1}^{V}\sum_{n\neq v}\ell_{sc}^{(vn)} + \sum_{v=1}^{V}\ell_{sa}^{(v)}. \tag{8}$$

3.5 Fine-grained dual contrastive mechanism

While feature-level contrast enables instance discrimination and coarse-grained semantic contrast ensures global cluster consistency, a significant information gap between these two levels often remains unaddressed by existing methods. To bridge this, we introduce a fine-grained dual contrastive mechanism, creating a 'semantic bridge' through an over-clustered semantic space. This mechanism explicitly aligns instance-level features with semantic groups via a cross-level contrast and ensures the bridge's structural integrity and discriminative power through a self-instance contrast. The following sections detail these two core components.

3.5.1 *Cross-level contrast mechanism*

We introduce a cross-level contrast mechanism to foster deep interaction between the feature-level contrastive and the fine-grained semantic cluster branch, enabling cross-level information fusion. Specifically, the mechanism

contrasts fine-grained cluster centers with high-level features extracted from the feature-level contrastive branch. This mechanism assigns each instance to the nearest centroid in the fine-grained semantic cluster branch, forming positive and negative pairs. A strong attraction exists between the instance and its corresponding cluster center in a positive pair, strengthening their association. Simultaneously, a repulsive force exists between positive and negative pairs, enhancing the distinction of negative pairs. To formalize this process, the latent representations $\{\mathbf{Z}^{(v)}\}_{v=1}^{V}$ are processed by a shared MLP to generate high-level features $\mathbf{G}^{(v)} \in \mathbb{R}^{N\times D_v}$, i.e., $\mathbf{G}^{(v)} = f_\theta(\mathbf{Z}^{(v)})$, where θ is the parameter of the shared MLP. We apply a clustering algorithm to perform over-clustering on these high-level features. After that, we compute C cluster centers in the fine-grained semantic cluster branch, denoted as $C^{(v)} = \{C_1^{(v)}; C_2^{(v)}; \ldots; C_2^{(V)}\}$. For each instance feature $\mathbf{h}_i^v$, we identify its corresponding cluster center $t^c(i)$. Then, $\mathbf{h}_i^v$ and its centroid $C_{t(i)}^{(v)}$ form a positive pair, while the rest cluster centers form the negative sample set $Neg^{(i)} \in \mathbb{R}^{(C-1)\times D_v}$ for the instance. Based on this setting, we define the cross-level contrast loss as

$$\ell_{cl}^{(vn)} = -\frac{1}{N}\sum_{i=1}^{C}\log\frac{e^{s(\mathbf{h}_i^v, C_{t(i)}^{(v)})/\tau_C}}{e^{s(\mathbf{h}_i^v, C_{t(i)}^{(v)})/\tau_C} + \sum_{j=1}^{C-1} e^{s(\mathbf{h}_i^v, Neg_j^{(i)})/\tau_C}}, \tag{9}$$

where τ_C is the cross-level contrast temperature parameter. By traversing all features, the cross-level contrast loss can be expressed as

$$\mathcal{L}_{CL} = \frac{1}{2}\sum_{v=1}^{V}\sum_{n\neq v}\ell_{cl}^{(vn)}. \tag{10}$$

The core intuition behind this cross-level contrast is to force the learned feature representation to be "aware" of its semantic destination. By pulling an instance's feature vector towards its corresponding cluster centroid from another view, we are not just comparing instance-to-instance; we are explicitly teaching the model to embed instances in a way that is inherently compatible with the semantic group structure. This novel instance-to-group contrast ensures that the feature learning process is directly guided by the clustering objective, facilitating a much more efficient information flow between the feature and semantic levels and enhancing the discriminative capability of the final representations.

3.5.2 *Self-instance contrast mechanism*

In addition to the cross-level contrast mechanism, we implement a detailed self-instance contrast mechanism within the fine-grained semantic clustering branch. To be specific, sample features from each view are based on another view. The index of the cluster center assigned to the representation of the i-th sample is defined as $u^c(i)$. Based on this setting, the contrast loss can be expressed as

$$\ell_{si}^{(vn)} = -\frac{1}{N}\sum_{i=1}^{C}\log\frac{e^{s(\mathbf{g}_i^v, C_{t(i)}^{(v)})/\tau_I}}{\sum_{j=1}^{C} e^{s(\mathbf{g}_i^v, \mathbf{g}_j^n)/\tau_I}}, \tag{11}$$

where τ_I is the self-instance contrast mechanism temperature parameter. By traversing all features, the self-instance contrast loss can be computed as

$$\mathcal{L}_{SI} = \frac{1}{2}\sum_{v=1}^{V}\sum_{n\neq v}\ell_{si}^{(vn)}. \tag{12}$$

The purpose of the self-instance contrast mechanism is to sharpen the definition and improve the quality of the fine-grained clusters themselves. By contrasting an instance g_i^v against all fine-grained centroids $C^{(v)}$ within its own view, we enforce a strong discriminative margin. This ensures that the fine-grained clusters are not just loose groupings but are compact and well-separated. A highly reliable fine-grained structure is essential, as it serves as the "anchor" for the cross-level contrast. In essence, this mechanism guarantees that the "semantic bridge" we are building is structurally sound, thereby maximizing the effectiveness of the entire dual-contrast framework.

3.6 Overall loss function

The objective function of CLMGC is thoughtfully designed, integrating multiple loss functions to optimize model performance comprehensively. These loss functions include feature-level contrastive loss $\mathcal{L}_{FC}$, which ensures feature space consistency. Coarse-grained semantic-level cluster contrast loss $\mathcal{L}_{SC}$ aims to achieve cluster consistency and ensure accuracy and reliability in clustering results. Cross-level contrast loss $\mathcal{L}_{CL}$ enhances interaction across levels,

enabling the model to better capture and utilize cross-level information associations. Fine-grained self-instance contrast loss $\mathcal{L}_{SI}$ focuses on strengthening discriminative learning between samples and their respective clusters within the fine-grained branch. Together, these loss functions enable CLMGC to achieve cross-level interaction and multi-granularity contrastive learning. The objective of CLMGC is defined as follows:

$$\mathcal{L} = \mathcal{L}_{FC} + \mathcal{L}_{SC} + \mathcal{L}_{CL} + \mathcal{L}_{SI}. \tag{13}$$

The training process of CLMGC is clearly explained in detail in Algorithm 1.

Algorithm 1 Training process of CLMGC.

Input: Dataset $\{X^{(v)}\}_{v=1}^{V}$; coarse-grained semantic cluster number K; fine-grained semantic cluster number C; temperature parameters τ_F, τ_S, τ_C, τ_I; training epochs E.

for epoch = 1 to E **do**
  Compute $\mathbf{Z}^{(v)}$ based on (1);
  Compute $\mathbf{H}^{(v)}$ based on (2);
  $\mathbf{Q}^{(v)} = f_\vartheta(\mathbf{Z}^{(v)})$;
  $\mathbf{G}^{(v)} = f_\theta(\mathbf{Z}^{(v)})$;
  Compute centroids in the fine-grained semantic cluster branch via K-Means;
  Compute feature-level contrastive loss $\mathcal{L}_{FC}$ based on (4) and (5);
  Compute coarse-grained semantic-level contrastive loss $\mathcal{L}_{SC}$ based on (6)–(8);
  Compute cross-level contrast loss $\mathcal{L}_{CL}$ based on (9) and (10);
  Compute self-instance contrast loss $\mathcal{L}_{SI}$ based on (11) and (12);
  Compute object loss $\mathcal{L}$ based on (13);
  Update model parameters to minimize $\mathcal{L}$;
end for

Output: Trained network.

4 Experiments

In this section, we introduce the model's implementation process. Afterward, we discuss the datasets, comparison methods, and evaluation metrics used in the experiments. Subsequently, we analyze the CLMGC model's clustering results and thoroughly discuss the experiment parameters. Finally, we comprehensively analyze the implementation of the ablation experiments.

4.1 Model implementation

The CLMGC training process includes multiple steps. Encoder extracts latent representations $\{\mathbf{Z}\}_{v=1}^{V}$, providing an initial abstraction of the data's intrinsic properties. Afterward, a feature Transformer and an MLP are used to refine high-level latent representations $\{\mathbf{H}\}_{v=1}^{V}$ and to determine cluster assignments $\{\mathbf{Q}\}_{v=1}^{V}$. This step enhances the model's understanding of the data's intrinsic structure. Building on this, we perform feature-level contrastive learning on high-level latent features $\{\mathbf{H}\}_{v=1}^{V}$. The core purpose of this step is to enhance the model's discriminative ability by minimizing similarities across instances, ensuring that each instance's uniqueness is emphasized. Simultaneously, we conduct a coarse-grained semantic comparison on cluster assignments $\{\mathbf{Q}\}_{v=1}^{V}$ to reduce label assignment errors caused by view-specific information, thus enhancing clustering accuracy and robustness. We introduce cross-level and self-instance contrast mechanisms to improve contrastive learning further and promote cross-level interaction and instance discrimination. These mechanisms enable the model to comprehensively capture and leverage information in contrastive learning, enhancing clustering performance. Finally, we apply the K-Means algorithm on refined high-level features $\{\mathbf{H}\}_{v=1}^{V}$ to obtain the final clustering results. We use the Adam optimizer during optimization and adjust the learning rate according to each dataset's characteristics. Expressly, learning rates were set to 5e−6, 5e−6, 5e−6, 5e−6, 1e−5, and 5e−5 for the DHA, ESP-Game, Flickr, NUS-Wide, COIL20, and MSRCv1 datasets, respectively, to optimize performance on each. This tuning process reflects our rigorous approach and commitment to optimizing model performance. All experiments were conducted on a computer with Ubuntu 20.04 LTS, featuring an Intel(R) Xeon(R) Gold-6133 CPU @ 2.50 GHz, NVIDIA TITAN XP GPU with 12 GB of memory.

4.2 Experiment setup

4.2.1 *Comparison baseline*

To thoroughly evaluate the effectiveness of our proposed CLMGC model, we selected seven state-of-the-art clustering methods for detailed comparative analysis. For single-view clustering, we selected the classic K-Means [33] algorithm

Table 1 Summary of baseline clustering methods.

Method	Category	Key contribution
K-Means [33]	Single-view, Traditional	A classic partitional clustering algorithm that serves as a fundamental benchmark.
MFLVC [30]	Deep MVC, Contrastive	Employs multi-level feature learning within a contrastive framework.
DSMVC [58]	Deep MVC	Aims to mitigate clustering performance degradation when adding more views.
DCIB [32]	Deep MVC, Information Bottleneck	Leverages the deep information bottleneck principle for cross-modal clustering.
DealMVC [48]	Deep MVC, Contrastive	Introduces a dual contrastive calibration mechanism to refine clustering.
MAGA [43]	Deep MVC, Contrastive	Integrates graph aggregation with confidence enhancement techniques.
SCMVC [28]	Deep MVC, Contrastive	Utilizes a self-weighted fusion strategy for contrastive learning results.
ConGMC [66]	Deep MVC, Information Theory	Maximizes mutual information between modalities under consistency guidance.

Table 2 Information of datasets.

Dataset	Samples	Views	Features	Clusters
DHA [69]	483	2	110/6144	23
ESP-Game [70]	11032	2	100/100	7
Flickr [71]	12154	2	100/100	6
NUS-Wide [72]	20000	2	100/100	8
COIL20 [67]	1440	3	1024/3304/6750	20
MSRCv1 [68]	210	5	24/576/512/256/254	7

as a benchmark due to its broad recognition and practicality, reporting optimal clustering results across all views. For MVC, we include MFLVC [30], DSMVC [58], DCIB [32], DealMVC [48], MAGA [43], SCMVC [28], and ConGMC [66] methods. The detailed information on the baseline methods is shown in Table 1. The comparative analysis with seven leading clustering methods aims to highlight the advantages and potential of the CLMGC model.

4.2.2 *Datasets*

To comprehensively evaluate the practical effectiveness of our model, we collected six widely used, publicly available multi-view datasets. These datasets are diverse in source and content, covering various sample types, including COIL20 [67], MSRCv1 [68], DHA [69], ESP-Game [70], Flickr [71], and NUS-Wide [72]. To provide a thorough understanding, we present critical details of these datasets in Table 2. The table includes the number of samples, views per sample, the feature dimensions of each view, and clusters for each dataset. This detailed dataset overview provides a comprehensive understanding, allowing us to assess model performance on these datasets more accurately.

4.2.3 *Evaluation metrics*

We employ seven standard clustering evaluation metrics: ACC, NMI, ARI, F-score, precision, recall, and purity. These metrics assess the consistency between clustering results and true labels, providing a robust basis for evaluating clustering performance. Notably, ARI and NMI values range from $[-1, 1]$, while ACC, F-score, precision, recall, and purity range from $[0, 1]$. Higher metric values indicate better clustering performance, reflecting a closer alignment between clustering results and true labels.

4.3 Clustering results

Table 3 presents the performance of CLMGC and various baseline methods across multi-view clustering tasks on six datasets. To thoroughly evaluate CLMGC's performance, we compare it with single-view clustering algorithms and multi-view clustering algorithms as a reference benchmark. These carefully designed comparative experiments aim to demonstrate CLMGC's superior performance and unique advantages in multi-view clustering from multiple perspectives. Specifically, we obtain the following observations.

Compared to the traditional single-view clustering method K-Means, which is widely used and relies on sample-based distance measures and a single feature space representation, it yields poor experimental results. This limitation is particularly evident with complex datasets such as Flickr and NUS-Wide, whose ACC and NMI are significantly lower than other methods. For instance, on the NUS-Wide dataset, the CLMGC model outperformed K-Means by approximately 25.17%, 20.58%, 22.89%, 19.85%, 14.04%, 16.39%, and 10.68% across multiple evaluation indicators. The main limitation of the K-Means method is its reliance on a single view for clustering, making it challenging to capture the multi-perspective information in the data fully. Consequently, the clustering performance is limited when the data and complex feature relationships are diverse.

Table 3 Clustering performance of various methods on six datasets. The highest and the second highest values under each metric are in bold and underlined, respectively. "–" indicates the presence of NaN values or memory overflow during model operation.

Method	DHA							ESP-Game						
	ACC (%)	NMI (%)	ARI (%)	Purity (%)	F-score (%)	Precision (%)	Recall (%)	ACC (%)	NMI (%)	ARI (%)	Purity (%)	F-score (%)	Precision (%)	Recall (%)
K-Means	64.39	76.63	49.34	65.42	61.73	58.76	65.02	50.72	34.95	25.20	58.79	38.15	38.53	37.79
MFLVC	51.76	69.98	42.28	52.38	52.09	41.53	69.85	42.80	25.10	19.65	49.14	32.10	30.67	33.67
DSMVC	60.46	75.61	51.23	62.11	59.60	55.79	63.98	30.91	17.32	11.52	41.17	24.58	24.48	24.68
DCIB	28.36	44.20	15.37	48.65	26.27	34.28	21.29	38.48	28.81	18.83	52.71	31.39	32.44	30.42
DealMVC	43.48	65.42	34.71	43.48	45.29	33.91	68.14	43.77	27.27	21.83	49.01	33.06	32.15	34.03
MAGA	31.26	56.88	23.58	31.68	34.26	23.79	61.19	37.17	19.53	13.61	44.37	26.51	26.75	26.27
SCMVC	71.43	77.45	57.39	72.46	63.17	62.35	64.01	46.81	31.76	25.17	53.58	35.93	35.33	36.54
ConGMC	29.19	62.29	22.28	30.23	41.28	26.80	**89.77**	42.85	25.49	21.73	48.82	36.28	28.59	**49.63**
Ours	**74.33**	**77.63**	**58.31**	**76.40**	**63.29**	**63.64**	62.95	**62.82**	**45.32**	**38.92**	**65.83**	**47.90**	**46.91**	48.93

Method	Flickr							NUS-Wide						
	ACC (%)	NMI (%)	ARI (%)	Purity (%)	F-score (%)	Precision (%)	Recall (%)	ACC (%)	NMI (%)	ARI (%)	Purity (%)	F-score (%)	Precision (%)	Recall (%)
K-Means	40.93	25.45	14.61	47.13	37.24	34.97	39.84	25.92	15.99	7.56	34.82	24.56	21.46	28.71
MFLVC	42.89	25.04	19.74	49.02	32.29	30.57	34.21	34.92	21.64	15.57	41.34	26.87	25.89	27.93
DSMVC	37.30	18.39	13.64	42.13	28.12	27.65	28.61	30.92	15.59	11.16	35.26	22.16	21.56	22.79
DCIB	35.82	26.17	17.49	52.72	34.49	34.91	34.08	30.54	22.37	13.64	39.62	25.68	25.30	26.06
DealMVC	43.77	27.27	21.83	49.01	33.06	32.15	34.03	36.33	24.31	17.96	44.63	28.15	27.78	28.54
MAGA	54.51	34.58	30.37	57.08	43.93	42.68	45.26	34.69	23.38	15.57	40.70	26.88	26.66	27.11
SCMVC	54.90	34.84	31.37	**56.84**	43.12	43.39	42.85	44.99	28.35	22.12	48.03	31.83	31.22	32.46
ConGMC	49.23	28.67	25.19	53.46	40.82	36.24	46.72	40.35	24.02	19.08	43.51	31.60	28.08	36.13
Ours	**55.95**	**39.78**	**35.13**	56.31	**45.44**	**45.10**	**45.79**	**51.09**	**36.57**	**30.45**	**54.67**	**38.60**	**37.85**	**39.39**

Method	COIL20							MSRCv1						
	ACC (%)	NMI (%)	ARI (%)	Purity (%)	F-score (%)	Precision (%)	Recall (%)	ACC (%)	NMI (%)	ARI (%)	Purity (%)	F-score (%)	Precision (%)	Recall (%)
K-Means	62.64	75.33	53.21	66.81	61.09	59.21	63.09	71.43	64.17	55.80	79.05	63.00	61.52	64.54
MFLVC	36.74	52.37	28.80	40.35	34.38	24.43	58.01	–	–	–	–	–	–	–
DSMVC	63.89	75.57	55.48	66.46	60.66	58.88	62.55	33.33	24.73	10.88	43.33	26.56	25.98	27.17
DCIB	33.68	44.73	20.34	47.43	32.75	34.33	31.31	47.62	44.78	29.13	60.48	43.27	47.14	39.98
DealMVC	28.47	49.68	23.30	32.29	30.01	19.51	**64.98**	43.81	41.95	26.88	44.29	45.30	33.37	70.51
MAGA	36.74	54.64	29.50	38.26	36.19	27.52	52.87	37.14	41.38	22.79	44.29	45.78	33.12	**74.13**
SCMVC	53.68	67.69	43.00	57.29	52.15	46.69	59.07	50.48	42.04	29.13	54.29	41.11	38.88	43.62
ConGMC	43.33	61.80	35.30	44.93	43.49	33.61	61.61	37.14	27.13	15.27	37.14	35.41	25.14	59.87
Ours	**69.51**	**77.04**	**58.69**	**71.67**	**64.58**	**62.78**	64.40	**83.81**	**72.57**	**66.72**	**86.67**	**73.08**	**72.15**	74.03

We compare seven deep MVC methods (MFLVC, DSMVC, DCIB, DealMVC, MAGA, SCMVC, and ConGMC). The MFLVC, DealMVC, MAGA, and SCMVC methods utilize multi-view contrastive learning models to enhance feature extraction by comparing information across different views. Despite this, most of these models are limited to direct inter-view comparisons, needing a comprehensive exploration of clustering structures in terms of cross-level and multi-granularity. In contrast, CLMGC incorporates a multi-granularity contrastive learning mechanism. This mechanism encompasses instance-level contrastive learning and extends to cluster-level contrastive learning, refining it into coarse-grained and fine-grained levels. This multi-level and multi-dimensional comparison strategy allows CLMGC to surpass multi-view contrastive learning methods like MFLVC, DealMVC, MAGA, and SCMVC in clustering performance, particularly in critical indicators such as ACC, NMI, and ARI. Notably, the advantages of CLMGC are even more pronounced when handling complex datasets like NUS-Wide and MSRCv1.

On the other hand, DCIB and ConGMC utilize the information bottleneck theory to extract representative features through information compression and selection. Although these methods perform well in specific scenarios, the information bottleneck approach may need help with multi-view data containing high feature redundancy, leading to suboptimal clustering results (e.g., lower F-score and Precision) on some datasets. In opposition, CLMGC achieves deep integration of fine-grained clustering features with instance-level features using its unique multi-level contrastive learning and cross-level interaction mechanism. This fusion enhances overall clustering performance and enables CLMGC to perform well in challenging scenarios for DCIB, such as the Flickr and ESP-Game datasets. Additionally, CLMGC incorporates an innovative self-instance contrast mechanism, further enhancing the compactness of fine-grained cluster-level features and ensuring strong performance across complex scenarios.

In summary, CLMGC demonstrates superior performance and broader application potential due to its advanced multi-view learning capability, cross-level semantic interaction mechanism, and refined fine-grained clustering strategy, outperforming single-view methods (e.g., K-Means) and leading multi-view clustering approaches (e.g., MFLVC, DSMVC, DCIB, DealMVC, MAGA, SCMVC, and ConGMC).

A comprehensive analysis of multiple evaluation metrics is crucial, as different metrics, such as ACC, NMI, and ARI, measure clustering quality from distinct perspectives, which explains the performance inconsistencies observed across different methods. For instance, some algorithms may excel at ensuring cluster completeness, leading to a high Recall or NMI, but at the cost of cluster purity, which results in lower Precision and can be heavily penalized by the strict, pair-based ARI. A clear example of this trade-off can be seen with the MAGA method on the MSRCv1 dataset, which achieves a high Recall (74.13%) but a much lower Precision (33.12%), indicating it creates large, impure clusters. In contrast, the superior performance of our proposed CLMGC method is demonstrated not just

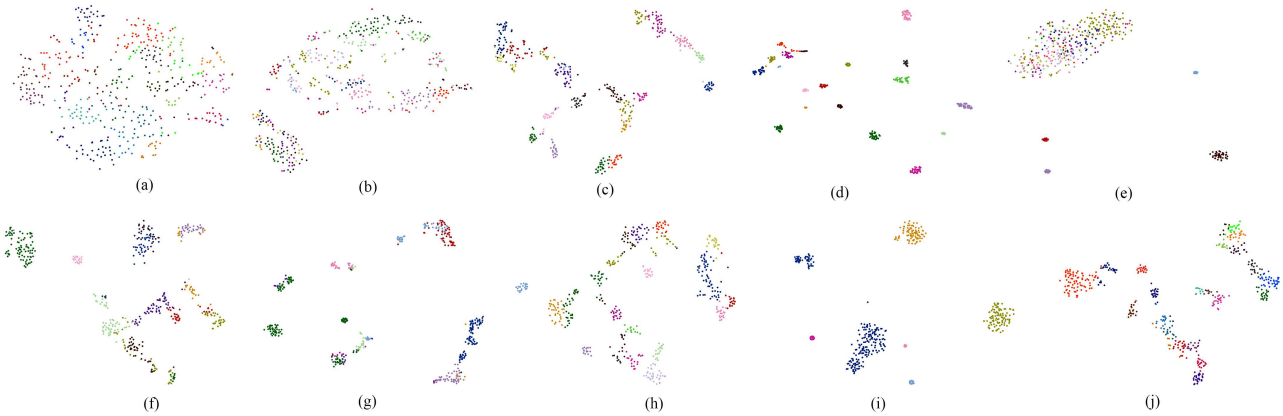

Figure 2 (Color online) Cluster visualization results. Visualization of features on the DHA dataset using different models. (a) Raw; (b) K-Means; (c) MFLVC; (d) DSMVC; (e) DCIB; (f) DealMVC; (g) MAGA; (h) SCMVC; (i) ConGMC; (j) Ours.

by its high scores but by its balance and consistency across all metrics. This is a direct consequence of its multi-granularity contrastive learning architecture. Specifically, the feature-level contrast and the fine-grained self-instance contrast work in tandem to maximize the discriminability between individual samples, forcing the model to learn tight and pure cluster boundaries, which directly boosts Precision and ARI. Simultaneously, the coarse-grained semantic contrast enforces global consistency across views, preventing the fragmentation of true classes and thus ensuring high Recall and NMI. It is this built-in mechanism of checks and balances that enables CLMGC, on the MSRCv1 dataset, to achieve both excellent Precision (72.15%) and Recall (74.03%), leading to a decisively superior F-score (73.08%) and ARI (66.72%). Therefore, this robust and well-rounded performance across diverse evaluation dimensions strongly validates that CLMGC learns a more fundamental and accurate data structure, rather than over-optimizing for any single aspect of clustering quality.

To intuitively highlight CLMGC's excellent clustering performance and capability in latent feature extraction, we design visualization experiments to demonstrate CLMGC's significant advantages. Specifically, we use the t-SNE algorithm [73] to visualize the embeddings learned by different models on the DHA datasets. As shown in Figure 2, a comparison of the original feature space, the other eight state-of-the-art models, and the embedding space generated by CLMGC reveals that CLMGC not only uncovers a more refined clustering structure but also demonstrates a significantly enhanced clustering effect, effectively proving its capability to explore data's intrinsic characteristics.

4.4 Parameter analysis

The CLMGC model demonstrates strong capability in fusing coarse-grained and fine-grained semantic information across levels. Specifically, when the coarse-grained clustering branch is set to match the final number of clustering targets, the fine-grained clustering branch typically uses a higher cluster count to achieve over-clustering. To thoroughly investigate the impact of the number of clusters C in the fine-grained clustering branch on model performance, we conducted experiments varying C from the target cluster number K to 400 and analyzed CLMGC's clustering effects under different C values. We select two datasets, COIL20 and MSRCv1, with target cluster numbers of 20 and 7, respectively. As shown in Figure 3, when C gradually increases from K to 200, CLMGC's performance metrics (ACC, NMI, ARI, Purity, F-score, Precision, and Recall) display a significant and rapid improvement. This performance improvement is primarily due to the richer and more detailed semantic information provided by over-clustering. When C increases from 200 to 400, CLMGC's clustering performance improves, though the gains are more diminutive. It is noteworthy that the curves, particularly for MSRCv1, do not reach a complete saturation point, which prompted our decision to balance further marginal improvements against practical considerations. Specifically, our choice to uniformly set C to 400 across all datasets is based on a deliberate trade-off. While a larger C might yield slight further gains, it would also incur a significant increase in computational cost. Furthermore, an excessively large C risks undermining the "grouping" aspect of the fine-grained semantic clusters. Therefore, we conclude that $C = 400$ represents a robust and efficient setting that allows the model to fully leverage fine-grained semantic information and achieve near-optimal performance in practical applications.

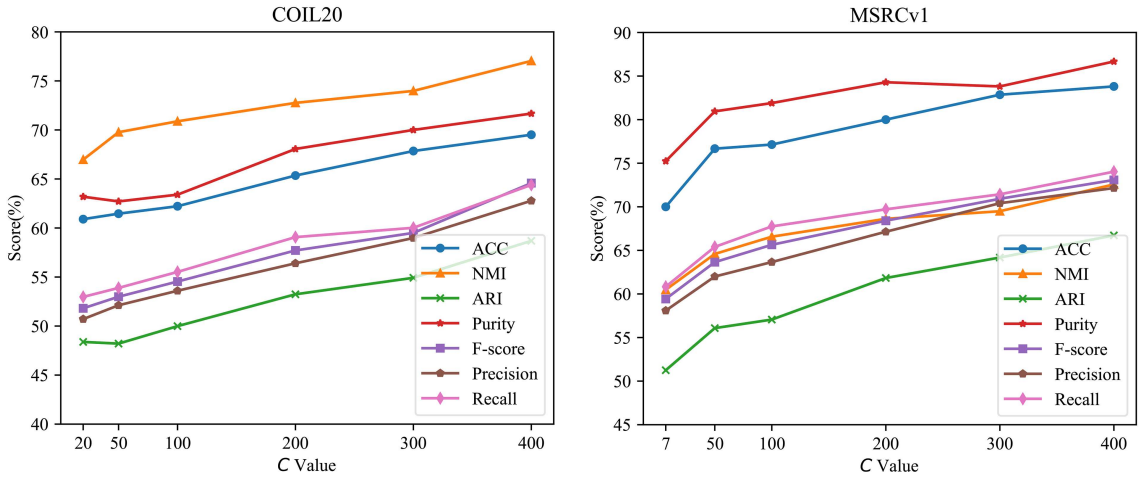

Figure 3 (Color online) Comparison of ACC, NMI, ARI, Purity, F-score, Precision, and Recall for different C values.

Table 4 The performance of CLMGC and its three variants on four evaluation metrics across six multi-view datasets.

Metrics				ACC (%)	NMI (%)	ARI (%)	Purity (%)	F-score (%)	Precision (%)	Recall (%)	ACC (%)	NMI (%)	ARI (%)	Purity (%)	F-score (%)	Precision (%)	Recall (%)
$\mathcal{L}_{FC}$	$\mathcal{L}_{SC}$	$\mathcal{L}_{CL}$	$\mathcal{L}_{SI}$	DHA							ESP-Game						
✓	✓	✓		71.01	76.90	56.94	75.98	61.56	61,31	61.81	54.25	35.21	29.13	55.78	40.18	39.10	41.32
✓	✓		✓	68.12	74.28	52.45	68.32	58.17	57.97	58.38	59.83	41.34	36.11	62.99	45.28	44.01	46.63
✓	✓			60.87	71.21	47.2	70.19	52.62	51.72	53.56	50.05	32.27	26.34	54.09	37.29	36.80	37.79
✓	✓	✓	✓	**74.33**	**77.63**	**58.31**	**76.40**	**63.29**	**63.64**	**62.95**	**62.82**	**45.32**	**38.92**	**65.83**	**47.90**	**46.91**	**48.93**
$\mathcal{L}_{FC}$	$\mathcal{L}_{SC}$	$\mathcal{L}_{CL}$	$\mathcal{L}_{SI}$	Flickr							NUS-Wide						
✓	✓	✓		53.83	35.79	30.72	56.92	43.06	43.71	42.43	48.79	33.28	26.60	52.25	36.45	36.09	36.81
✓	✓		✓	53.13	36.10	30.45	55.81	42.95	43.63	42.29	50.27	35.77	30.30	55.65	38.58	37.62	39.60
✓	✓			51.06	31.46	28.70	54.89	40.52	40.45	40.59	44.57	29.39	22.12	50.13	32.19	32.21	32.17
✓	✓	✓	✓	**55.95**	**39.78**	**35.13**	**56.31**	**45.44**	**45.10**	**45.79**	**51.09**	**36.57**	**30.45**	**54.67**	**38.60**	**37.85**	**39.39**
$\mathcal{L}_{FC}$	$\mathcal{L}_{SC}$	$\mathcal{L}_{CL}$	$\mathcal{L}_{SI}$	COIL20							MSRCv1						
✓	✓	✓		68.47	76.07	57.88	70.42	62.71	61.54	63.93	82.38	69.95	63.61	82.86	70.89	70.21	71.59
✓	✓		✓	67.15	75.74	56.94	68.96	61.65	60.44	62.92	81.90	68.88	62.51	82.86	69.80	69.28	70.32
✓	✓			63.40	73.12	53.79	66.6	58.93	56.68	61.37	79.05	64.29	57.16	82.38	65.71	65.14	66.29
✓	✓	✓	✓	**69.51**	**77.04**	**58.69**	**71.67**	**64.58**	**62.78**	**64.40**	**83.81**	**72.57**	**66.72**	**86.67**	**73.08**	**72.15**	**74.03**

4.5 Ablation studies

Based on [30], to empirically justify our core architectural choices and validate the effectiveness of the proposed multi-granularity contrastive framework, we conducted a series of ablation experiments. Specifically, we analyzed the two core contrast mechanisms in the CLMGC model: cross-level contrast $\mathcal{L}_{CL}$ loss and self-instance contrast loss $\mathcal{L}_{SI}$. We focused on three model variants to analyze their impact on clustering performance comprehensively. (1) Remove self-instance contrast loss $\mathcal{L}_{SI}$: This variant explores the model's clustering performance without the self-instance contrastive mechanism to see if clustering ability weakens. (2) Remove cross-level contrast loss $\mathcal{L}_{CL}$: This variant evaluates the role of the cross-level contrast mechanism in facilitating information flow and fusion across views and its contribution to clustering accuracy. (3) Remove both self-instance and cross-level contrast losses: This extreme case helps us understand the baseline clustering performance of CLMGC without any contrast mechanisms, which is used to validate the role of the introduced fine-grained branch in cross-level semantic interaction. To comprehensively and objectively assess these model variants, we employ seven clustering evaluation metrics and perform exhaustive testing on six real-world datasets. As detailed in Table 4, we comprehensively evaluated three variants of our CLMGC model across six datasets. The analysis below not only demonstrates the contribution of each component but also provides a direct experimental validation for the rationale behind our two-stage coarse- and fine-grained clustering design.

The necessity of the fine-grained semantic bridge: The most critical comparison is our full model against a variant where the entire fine-grained dual contrastive mechanism ($\mathcal{L}_{CL}$ and $\mathcal{L}_{SI}$) is removed. This variant represents a more conventional architecture that combines only instance-level contrast with a single, coarse-grained semantic objective. The results show a dramatic and consistent drop in performance. Using the DHA dataset as an example, this removal causes the ACC to plummet by 13.46% (from 74.33% to 60.87%) and the ARI to drop by 11.11% (from 58.31% to 47.2%). This severe degradation is not merely a performance dip; it is direct experimental proof that the fine-grained semantic level is the core engine of our model's success. It acts as an essential "semantic bridge" that enables effective, deep interaction between feature representations and group structures, a capability that a single

Table 5 Total running time (in seconds) on different datasets.

Dataset	MFLVC	DSMVC	DCIB	DealMVC	MAGA	SCMVC	ConGMC	Ours
DHA	36.93	10.62	56.64	37.6	253.22	104.47	53.16	112.91
ESP-Game	246.08	148.99	131.31	282.56	9315.44	909.07	1073.57	626.09
Flickr	267.52	165.14	139.67	304.29	10126.54	940.67	1176.78	661,71
NUS-Wide	506.52	268.92	236.72	565.83	14870.97	1194.55	2008.56	1061.98
COIL20	155.23	32.24	63.96	95.99	1285.97	611.79	149.58	285.07
MSRCv1	35.17	4.69	25.7	23.31	212.86	97.61	14.23	85.96

coarse-grained objective alone cannot provide.

The synergistic roles of the dual contrast mechanisms: Having established the necessity of the fine-grained level, we now analyze its internal components to understand how the "bridge" functions. When only the cross-level contrast loss ($\mathcal{L}_{CL}$) is removed, performance still declines significantly. As demonstrated by the MSRCv1 dataset, ACC drops by 1.91% and ARI by 4.21%. This confirms the crucial role of $\mathcal{L}_{CL}$ in actively connecting the feature and semantic levels. It is the mechanism that facilitates deep information transfer across the bridge, ensuring that feature learning is directly guided by the fine-grained semantic structure. Similarly, removing only the self-instance contrast loss ($\mathcal{L}_{SI}$) also leads to a notable performance decline. On the ESP-Game dataset, ACC drops by 8.57% and ARI by 9.79%. This validates the role of $\mathcal{L}_{SI}$ in ensuring the structural integrity of the bridge itself. Enhancing the discriminative power within the fine-grained branches makes the semantic anchors (the fine-grained centroids) more compact and well-separated, thus making the cross-level interaction more stable and meaningful.

In conclusion, our comprehensive ablation study provides a robust empirical justification for our design choices. The results clearly show that the fine-grained level is indispensable for achieving high performance, and that the coarse-grained level (retained in all variants) provides the necessary global guidance. The synergy between these two stages, powered by the dual contrast mechanisms, is the fundamental reason for CLMGC's superior performance.

4.6 Model complexity analysis

To understand CLMGC's practical applicability, we first theoretically analyze its computational cost. The model's complexity stems from three main components: view-specific feature extraction (encoders and MLP projectors) with $\mathcal{O}(N\sum_v(D_v d))$, global feature enhancement and contrast (Transformer and cross-view contrast) incurring $\mathcal{O}(VN^2d)$, and the multi-granularity semantic learning module (fine-grained K-Means and semantic-level contrast losses) with $\mathcal{O}(VNCd)$. Here, N is the number of samples, D_v is the input dimension for view v, d is the latent space dimension, V is the number of views, and C is the number of fine-grained clusters. Given that N^2 is the dominant term, the overall time complexity of CLMGC is $\mathcal{O}(VN^2d)$.

For practical comparison, we evaluated CLMGC's training time against seven representative deep MVC methods on six datasets, as summarized in Table 5. Our proposed model exhibits a competitive running time profile. Although not the fastest among lightweight models, CLMGC demonstrates significant efficiency over several computationally demanding state-of-the-art methods, particularly on large-scale datasets like NUS-Wide, where it is nearly twice as fast as ConGMC and over 10 times faster than MAGA. This controlled increase in complexity, balanced against the substantial clustering accuracy improvements detailed in Table 3, positions CLMGC as a practical and effective solution achieving a strong performance-efficiency trade-off.

5 Conclusion

In this paper, we propose an innovative cross-level interactive and multi-granularity contrastive learning method, named CLMGC, to optimize the MVC task. Unlike previous MVC methods limited to feature-level or/and semantic-level contrastive learning, CLMGC introduces the fine-grained cluster at the director level and incorporates a sophisticated dual contrast mechanism. Specifically, we perform contrastive learning at both the feature and coarse-grained semantic levels to explore the correlation between local feature contrast and global semantic information. Meanwhile, we also implement a fine-grained dual contrastive mechanism to explore complex hybrid granularity connections between feature and semantic branches. We achieve more accurate and reliable clustering results by combining feature-level contrastive learning, semantic-level contrastive learning, and the dual contrastive mechanism. Experiments conducted on six multi-view datasets demonstrate that the CLMGC model achieves superior performance compared to the latest deep multi-view clustering methods. In the future, we will focus on establishing a bidirectional information flow across the semantic hierarchy. Furthermore, we aim to leverage this more robust

and adaptive structure to address challenging real-world scenarios, particularly extending our model to handle incomplete multi-view clustering.

Acknowledgements The work was supported in part by National Natural Science Foundation of China (Grant Nos. 62476258, 62522604), Natural Science Foundation of Hubei Province (Grant No. 2025AFA113), and Fundamental Research Funds for National Universities, China University of Geosciences (Wuhan) (Grant No. 2024XLB6).

References

1 Luo S, Zhang C, Zhang W, et al. Consistent and specific multi-view subspace clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018
2 Wang Y, Chang D, Fu Z, et al. Consistent multiple graph embedding for multi-view clustering. IEEE Trans Multimedia, 2021, 25: 1008–1018
3 Liang J Y, Liu X L, Bai L, et al. Incomplete multi-view clustering via local and global co-regularization. Sci China Inf Sci, 2022, 65: 152105
4 Fang U, Li M, Li J, et al. A comprehensive survey on multi-view clustering. IEEE Trans Knowl Data Eng, 2023, 35: 12350–12368
5 Chen Y, Zhang X, Wang J, et al. Large-scale multi-view clustering based on anchor strategy and tensor collaborative learning. In: Proceedings of International Conference on Cloud Computing and Intelligent Systems, 2023. 18–23
6 Chen Y, Wang S, Xiao X, et al. Self-paced enhanced low-rank tensor kernelized multi-view subspace clustering. IEEE Trans Multimedia, 2021, 24: 4054–4066
7 Qin Y, Pu N, Wu H. Elastic multi-view subspace clustering with pairwise and high-order correlations. IEEE Trans Knowledge Data Eng, 2024, 36: 556–568
8 Liu B Y, Huang L, Wang C D, et al. Multi-view consensus proximity learning for clustering. IEEE Trans Knowledge Data Eng, 2022, 34: 3405–3417
9 Li R, Zhang C, Fu H, et al. Reciprocal multi-layer subspace learning for multi-view clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 8172–8180
10 Li Z, Tang C, Zheng X, et al. High-order correlation preserved incomplete multi-view subspace clustering. IEEE Trans Image Process, 2022, 31: 2067–2080
11 Zou X, Tang C, Zheng X, et al. Inclusivity induced adaptive graph learning for multi-view clustering. Knowledge-Based Syst, 2023, 267: 110424
12 Zhang P, Liu X, Xiong J, et al. Consensus one-step multi-view subspace clustering. IEEE Trans Knowl Data Eng, 2022, 34: 4676–4689
13 Qin Y, Pu N, Wu H. EDMC: efficient multi-view clustering via cluster and instance space learning. IEEE Trans Multimedia, 2024, 26: 5273–5283
14 Wen J, Zhang Z, Xu Y, et al. Incomplete multi-view clustering via graph regularized matrix factorization. In: Proceedings of the European Conference on Computer Vision, 2018
15 Yang Z, Liang N, Yan W, et al. Uniform distribution non-negative matrix factorization for multiview clustering. IEEE Trans Cybern, 2020, 51: 3249–3262
16 Li Z, Tang C, Liu X, et al. Tensor-based multi-view block-diagonal structure diffusion for clustering incomplete multi-view data. In: Proceedings of IEEE International Conference on Multimedia and Expo, 2021. 1–6
17 Liu S, Liao Q, Wang S, et al. Robust and consistent anchor graph learning for multi-view clustering. IEEE Trans Knowl Data Eng, 2024, 36: 4207–4219
18 Nie F, Li J, Li X, et al. Self-weighted multiview clustering with multiple graphs. In: Proceedings of IJCAI, 2017. 2564–2570
19 Li Z, Tang C, Liu X, et al. Consensus graph learning for multi-view clustering. IEEE Trans Multimedia, 2021, 24: 2461–2472
20 Li Z, Tang C, Zheng X, et al. Mutual structure learning for multiple kernel clustering. Inf Sci, 2023, 647: 119445
21 Li L, He H. Bipartite graph based multi-view clustering. IEEE Trans Knowledge Data Eng, 2022, 34: 3111–3125
22 Xia W, Wang Q, Gao Q, et al. Self-supervised graph convolutional network for multi-view clustering. IEEE Trans Multimedia, 2022, 24: 3182–3192
23 Fu S C, Peng Q M, He Y E, et al. Unsupervised multiplex graph diffusion networks with multi-level canonical correlation analysis for multiplex graph representation learning. Sci China Inf Sci, 2025, 68: 132102
24 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444
25 Alwassel H, Mahajan D, Korbar B, et al. Self-supervised learning by cross-modal audio-video clustering. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 33: 9758–9770
26 Tang C, Wang J, Zheng X, et al. Spatial and spectral structure preserved self-representation for unsupervised hyperspectral band selection. IEEE Trans Geosci Remote Sensing, 2023, 61: 1–13
27 Yan W, Zhang Y, Lv C, et al. Gcfagg: global and cross-view feature aggregation for multi-view clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 19863–19872
28 Wu S, Zheng Y, Ren Y, et al. Self-weighted contrastive fusion for deep multi-view clustering. IEEE Trans Multimedia, 2024, 26: 9150–9162
29 Yan W, Zhou Y, Wang Y, et al. Multi-view semantic consistency based information bottleneck for clustering. Knowledge-Based Syst, 2024, 288: 111448
30 Xu J, Tang H, Ren Y, et al. Multi-level feature learning for contrastive multi-view clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 16051–16060
31 Xu J, Ren Y, Tang H, et al. Self-supervised discriminative feature learning for deep multi-view clustering. IEEE Trans Knowl Data Eng, 2022, 35: 7470–7482
32 Yan X, Mao Y, Ye Y, et al. Cross-modal clustering with deep correlated information bottleneck method. IEEE Trans Neural Netw Learn Syst, 2023, 35: 13508–13522
33 Hartigan J A, Wong M A. Algorithm as 136: a k-means clustering algorithm. J Royal Stat Soc Ser C, 1979, 28: 100–108
34 von Luxburg U. A tutorial on spectral clustering. Stat Comput, 2007, 17: 395–416
35 Wei S, Wang J, Yu G, et al. Deep incomplete multi-view multiple clusterings. In: Proceedings of IEEE International Conference on Data Mining, 2020. 651–660
36 Huang Z, Ren Y, Pu X, et al. Non-linear fusion for self-paced multi-view clustering. In: Proceedings of the ACM International Conference on Multimedia, 2021. 3211–3219
37 Xiao Y, Yang D, Li J, et al. Dual alignment feature embedding network for multi-omics data clustering. Knowledge-Based Syst, 2025, 309: 112774
38 Zhu Y, He X, Tang C, et al. Multi-view adaptive fusion network for spatially resolved transcriptomics data clustering. IEEE Trans Knowl Data Eng, 2024, 36: 8889–8900
39 Deng S, Zheng X, Tang C, et al. scspaf: cell similarity purified adaptive fusion network for single-cell multi-omics clustering. IEEE Trans Comput Biol Bioinform, 2025, doi: 10.1109/TCBBIO.2025.3608251
40 Lu Y, Li Q, Zhang X, et al. Deep contrastive representation learning for multi-modal clustering. Neurocomputing, 2024, 581: 127523
41 Mao Y, Yan X, Hu S, et al. Contrastive cross-modal clustering with twin network. Pattern Recogn, 2024, 155: 110645
42 Zou G, Ye Y, Chen T, et al. Learning dual enhanced representation for contrastive multi-view clustering. In: Proceedings of the ACM International Conference on Multimedia, 2024. 8731–8739

43 Bian J, Xie X, Lai J H, et al. Multi-view contrastive clustering via integrating graph aggregation and confidence enhancement. Inf Fusion, 2024, 108: 102393
44 Hu S, Zou G, Zhang C, et al. Joint contrastive triple-learning for deep multi-view clustering. Inform Processing Manage, 2023, 60: 103284
45 Chen J, Mao H, Woo W L, et al. Deep multiview clustering by contrasting cluster assignments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 16752–16761
46 Hu S, Zhang C, Zou G, et al. Deep multiview clustering by pseudo-label guided contrastive learning and dual correlation learning. IEEE Trans Neural Networks Learn Syst, 2025, 36: 3646–3658
47 Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: Proceedings of International Conference on Machine Learning, 2020. 1597–1607
48 Yang X, Jiaqi J, Wang S, et al. Dealmvc: dual contrastive calibration for multi-view clustering. In: Proceedings of the ACM International Conference on Multimedia, 2023. 337–346
49 Li Y, Hu P, Liu Z, et al. Contrastive clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 8547–8555
50 Zhang C, Fu H, Hu Q, et al. Generalized latent multi-view subspace clustering. IEEE Trans Pattern Anal Mach Intell, 2018, 42: 86–99
51 Liu S, Wang S, Zhang P, et al. Efficient one-pass multi-view subspace clustering with consensus anchors. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2022. 7576–7584
52 Jia Y, Liu H, Hou J, et al. Multi-view spectral clustering tailored tensor low-rank representation. IEEE Trans Circuits Syst Video Technol, 2021, 31: 4784–4797
53 Yang W, Tang C, Zheng X, et al. Eigenvalue ratio inspired partition learning and fusion for multiple kernel clustering. IEEE Trans Knowl Data Eng, 2024, 36: 8135–8147
54 Yang W, Tang C, Liu X, et al. Smooth multiple kernel k-means via underlying graph filtering. IEEE Trans Neural Netw Learn Syst, 2025, 36: 14855–14868
55 Wei S, Wang J, Yu G, et al. Multi-view multiple clusterings using deep matrix factorization. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 6348–6355
56 Tang C, Zheng X, Zhang W, et al. Unsupervised feature selection via multiple graph fusion and feature weight learning. Sci China Inf Sci, 2023, 66: 152101
57 Fan S, Wang X, Shi C, et al. One2multi graph autoencoder for multi-view graph clustering. In: Proceedings of the Web Conference, 2020. 3070–3076
58 Tang H, Liu Y. Deep safe multi-view clustering: Reducing the risk of clustering performance degradation caused by view increase. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 202–211
59 Xiao Y, Tang C, Zheng X, et al. Mutual calibration network for multi-view clustering. IEEE Trans Circuits Syst Video Technol, 2025, doi: 10.1109/TCSVT.2025.3588889
60 Deng S, Zheng X, Tang C, et al. Find true collaborators: Banzhaf index-based cross view alignment for partially view-aligned clustering. In: Proceedings of the ACM International Conference on Multimedia, 2025
61 Zhang X, Qiu H, Liang W, et al. Generalization performance of ensemble clustering: From theory to algorithm. In: Proceedings of the International Conference on Machine Learning, 2025. 1–27
62 Wang T, Isola P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: Proceedings of International Conference on Machine Learning, 2020. 9929–9939
63 La Rosa L E C, Oliveira D A B. Learning from label proportions with prototypical contrastive clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2022. 2153–2161
64 Li H, Zhang L, Su K. Dual mutual information constraints for discriminative clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2023. 8571–8579
65 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 30
66 Mao Y, Yan X, Liu J, et al. ConGMC: consistency-guided multimodal clustering via mutual information maximin. IEEE Trans Multimedia, 2023, 26: 5131–5146
67 Nene S A, Nayar S K, Murase H, et al. Columbia Object Image Library (COIL-20). Technical Report CUCS-005-96, 1996
68 Zhao X, Niu X, Ma Y, et al. A multi-view ensemble clustering approach using joint entropy. Expert Syst Appl, 2024, 255: 124683
69 Lin Y C, Hu M C, Cheng W H, et al. Human action recognition and retrieval using sole depth information. In: Proceedings of the ACM International Conference on Multimedia, 2012. 1053–1056
70 von Ahn L, Dabbish L. Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2004. 319–326
71 Huiskes M J, Lew M S. The Mir Flickr retrieval evaluation. In: Proceedings of the ACM International Conference on Multimedia Information Retrieval, 2008. 39–43
72 Chua T S, Tang J, Hong R, et al. 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, 2009. 1–9
73 van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res, 2008, 9: 2579–2605