Mixture of Clustering Experts with Dual Consistency for Multi-View Clustering

Daidai Zhu[1], **Yang Zhao**[1,*], **Dandan Ma**[1], **Ganchao Liu**[1] and **Zhiyu Jiang**[1]

[1] School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China

daidaizhu@mail.nwpu.edu.cn, madandan@nwpu.edu.cn, {zhaoyang.opt, ganchliu, zhuyu.jiang.chn}@gmail.com

Abstract

Multi-view clustering aims to discover shared semantic structures from multiple complementary data views to improve clustering performance. However, most existing methods rely on a single clustering representation for each semantic cluster, which limits the ability to model complex cluster structures and diverse patterns in real-world data. To address this limitation, we propose a Mixture of Clustering Experts with Dual Consistency (MoCE-DC). MoCE-DC introduces a set of learnable clustering experts that characterize the same semantic cluster from multiple perspectives within a unified prototype space. A gating mechanism is employed to adaptively select experts for each sample. MoCE-DC decomposes complex semantic clusters into multiple collaborative sub-structures, thereby significantly enhancing the ability to model intricate intra-cluster diversity. To further ensure cross-view semantic consistency, we propose a dual-level alignment mechanism that enforces prediction consistency across views while guiding clustering assignments toward a more discriminative direction. In addition, a gating balance regularization strategy is introduced to mitigate expert collapse and promote balanced expert utilization. Extensive experiments on multiple public multi-view clustering benchmarks demonstrate that MoCE-DC outperforms state-of-the-art methods.

1 Introduction

Multi-view clustering [Chao *et al.*, 2021; Fang *et al.*, 2023b; Zhou *et al.*, 2024] aims to uncover latent semantic structures by integrating complementary information from multiple heterogeneous data sources. Compared to single-view approaches, it can more fully exploit both the consistency and complementarity across different views and achieve more accurate clustering [Wang *et al.*, 2023], with wide applications in computer vision, multimedia analysis, and information retrieval.

*Corresponding author

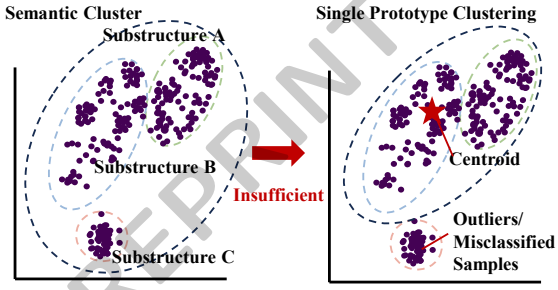


Figure 1: The left figure illustrates that a semantic cluster often comprises multiple latent substructures, exhibiting significant internal diversity. A clustering approach centered on a single prototype struggles to capture complex substructure distributions, as it relies solely on a single prototype for classification. This approach is prone to generating outlier samples or misclassified samples.

In recent years, multi-view clustering has achieved significant progress owing to the powerful nonlinear representation learning capability of deep learning. Autoencoder-based methods [Yang *et al.*, 2021; Xu *et al.*, 2021; Yan *et al.*, 2023; Bai *et al.*, 2024] learn latent representations from multi-view data and integrate cross-view information through feature fusion mechanisms. Graph-based approaches [Wang *et al.*, 2021; Huang *et al.*, 2023; Ling *et al.*, 2023] further explicitly model the relational structure among samples by fusing multi-view graph structures to capture cross-view similarities, thereby enhancing clustering discriminability. In addition, contrastive learning–based methods [Ke *et al.*, 2021; Pan and Kang, 2021; Feng *et al.*, 2024] design cross-view contrastive objectives to enforce representation consistency of semantically identical samples across different views, facilitating the learning of discriminative features. For example, [Zhang *et al.*, 2019] designed a nested autoencoder framework to perform view-specific representation learning and multi-view information encoding within a unified framework. [Feng *et al.*, 2024] introduced view-specific contrastive encoders and generated pseudo labels on fused representations, utilizing these pseudo labels to guide the positive-negative pair selection in contrastive encoders. By effectively exploiting both the consistency and complementarity among views, these deep multi-view clustering methods demonstrate remarkable clustering performance.

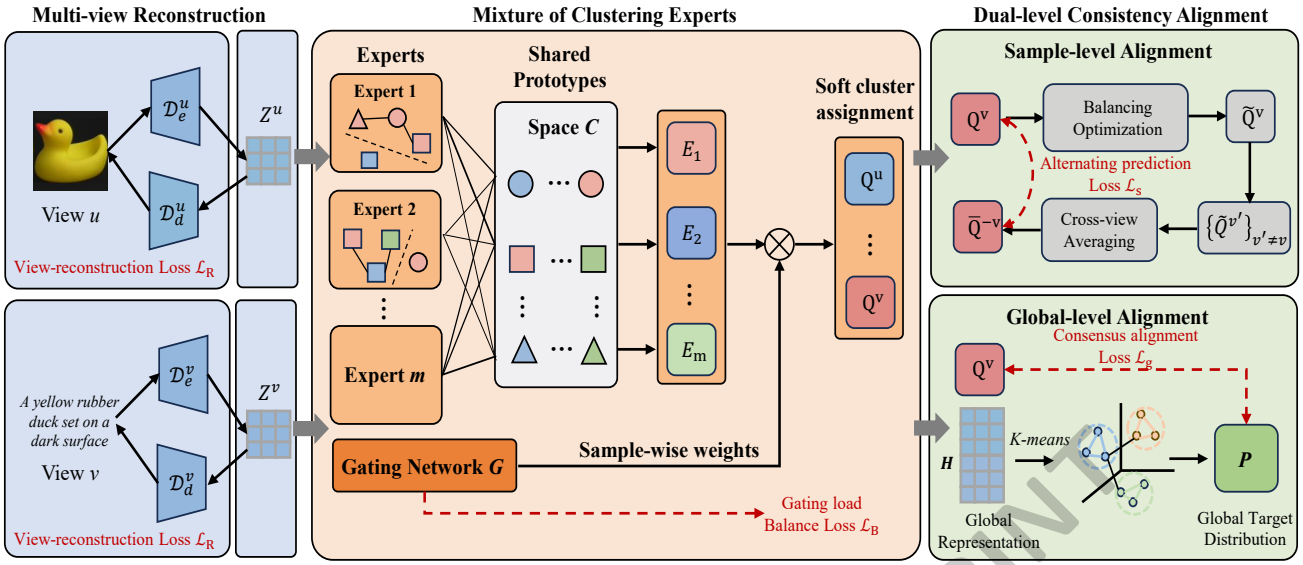


Figure 2: The overall framework of MoCE-DC. The framework consists of three components. First, multi-view encoders are employed to learn view-specific representations for clustering. Second, a mixture of clustering experts module captures intra-cluster substructures by leveraging multiple experts. The outputs of the experts are adaptively weighted by a gating network to generate view-specific cluster assignments. Third, a dual-level consistency alignment module is introduced, where sample-level cross-view prediction enforces local alignment and a global target distribution guides all views toward a unified clustering structure.

However, most existing multi-view clustering methods characterize each semantic cluster using a single prototype and assume a compact cluster assignment. Some representative approaches [Xu *et al.*, 2022b; Chen *et al.*, 2023; Yang *et al.*, 2021; Yan *et al.*, 2025] introduce an explicit clustering layer on top of learned representations and perform end-to-end optimization. Although such designs are structurally simple and efficient, they essentially assume that each semantic cluster can be adequately represented by a single centroid. However, data distributions often exhibit diversity within clusters. As illustrated in Figure 1, samples belonging to the same semantic category may contain multiple latent substructures. Modeling such complex distributions with a single prototype is therefore inadequate to capture the underlying intra-cluster substructure, which inevitably limits clustering performance. To alleviate this limitation, several graph-based methods [Zhang *et al.*, 2024; Chen *et al.*, 2025] attempt to enhance intra-cluster modeling by constructing sample-to-anchor graphs. Nevertheless, their performance heavily depends on the quality of graph construction, which is often sensitive to noise. Moreover, graph structures are usually built independently for each view, making it difficult to ensure semantic consistency across views. Therefore, how to effectively model complex intra-cluster structures while simultaneously preserving semantic consistency across multiple views remains a challenge.

To address the challenge, we propose a Mixture of Clustering Experts with Dual Consistency for Multi-View Clustering termed MoCE-DC. The framework is shown in Figure 2. Specifically, MoCE-DC introduces a set of learnable clustering experts to collaboratively model complex cluster structures. Each expert captures complementary semantic characteristics within a shared prototype space, enabling the model to decompose intricate clusters into multiple cooperative substructures. To facilitate adaptive expert utilization, a gating mechanism is employed to dynamically weight expert contributions according to view-specific representations, allowing each sample to be clustered through the most relevant expert combinations. To further ensure consistency across views, MoCE-DC incorporates a dual-level consistency alignment mechanism. At the sample level, cross-view prediction consistency is enforced by aligning the cluster assignment of each sample with the consensus of its assignments in the other views. At the global level, a consensus target distribution is constructed by aggregating multi-view information, which serves as a unified semantic reference to guide all views toward a consensus clustering structure and effectively mitigate semantic drift. In addition, to alleviate the commonly observed expert collapse issue in mixture-of-experts models, we introduce a gating balance regularization that promotes balanced expert activation and encourages cooperative expert learning. Collectively, these components enable MoCE-DC to simultaneously refine intra-cluster substructure modeling and enforce cross-view semantic consistency.

The main contributions of this work can be summarized as follows:

1. We propose a novel multi-view clustering framework, termed MoCE-DC, which explicitly models intra-cluster diversity by introducing multiple cooperative clustering experts. MoCE-DC enables different experts to capture complementary substructures within the same semantic cluster, significantly enhancing the expressive capacity for complex clusters.
2. We propose a dual-consistency alignment strategy to en-

hance cross-view semantic consistency. By jointly enforcing sample-level prediction consistency and global semantic agreement across views, MoCE-DC effectively reduces cross-view semantic discrepancy and stabilizes the clustering process.

3. Extensive experimental results on several benchmark datasets demonstrate that the proposed method consistently outperforms existing state-of-the-art multi-view clustering approaches.

2 Method

In this section, we introduce our proposed method MoCE-DC. First, we introduce the key notations employed throughout this paper. Consider a multi-view dataset denoted as $\{\mathbf{X}^v \in \mathbb{R}^{N\times d_v}\}_{v=1}^V$, where N denotes the number of samples, V represents the number of views, d_v denotes the sample dimension of the v-th view, and $\mathbf{x}_i^v$ denotes the v-th view of the i-th sample. The goal of multi-view clustering is to partition the N samples into C clusters by effectively fusing information from all V views.

To achieve this, we first project the multi-view data into a low-dimensional latent space using a set of view-specific deep autoencoders. Formally, for the v-th view, we construct an encoder network $\mathcal{D}_e^v$ and a decoder network $\mathcal{D}_d^v$, parameterized by θ^v and ϕ^v, respectively. Given an input sample $\mathbf{x}_i^v$, its latent representation is derived as $\mathbf{z}_i^v = \mathcal{D}_e^v(\mathbf{x}_i^v;\theta^v)$. These autoencoders are trained by minimizing a reconstruction loss, which ensures that the essential information of each view is preserved in the latent space:

$$\mathcal{L}_R = \sum_{v=1}^{V}\sum_{i=1}^{N} \|\mathbf{x}_i^v - \mathcal{D}_d^v(\mathbf{z}_i^v;\phi^v)\|_2^2. \tag{1}$$

2.1 Mixture of Clustering Experts

To explicitly model the inherent diversity and latent substructures within clusters, we introduce a Mixture of Clustering Experts module (MoCE). This module associates each cluster with a set of specialized experts, with each expert dedicated to capturing a latent substructure present within the cluster. Concurrently, we design a learnable gating network that dynamically routes samples to relevant experts, enabling the model to adaptively compose complex clusters from these elementary substructures.

Formally, given the view-specific latent representations $\{\mathbf{Z}^v\}_{v=1}^V$, the MoCE module generates view-specific soft cluster assignments $\mathbf{Q}^v \in \mathbb{R}^{N\times C}$ by adaptively aggregating the predictions of M experts:

$$\mathbf{Q}^v = \sum_{m=1}^{M} \mathbf{G}_m^v \odot \mathbf{E}_m^v, \tag{2}$$

where $\mathbf{E}_m^v \in \mathbb{R}^{N\times C}$ denotes the soft cluster assignment produced by the m-th expert for view v, and $\mathbf{G}_m^v \in \mathbb{R}^{N\times 1}$ represents the corresponding sample-wise gating weights. The Hadamard product $\odot$ applies the gating weights to expert predictions in a sample-wise manner. Under this formulation, samples belonging to the same cluster can be jointly explained by several experts with varying activation strengths, enabling MoCE to capture complex cluster geometries.

Shared-Prototype-Based Clustering Experts. To explicitly model diverse clustering behaviors while preserving cross-view consistency, MoCE introduces a set of shared clustering prototypes $\{\boldsymbol{\mu}_k\}_{k=1}^C$, which are shared across all experts and views. These prototypes serve as common semantic anchors, enabling different experts to produce distinct clustering patterns while remaining aligned in a unified prototype space.

Given a sample i from view v, the m-th expert first transforms the representation via an expert-specific function $f_m(\cdot)$ and then computes its soft assignment to each prototype. The Student's t-distribution is adopted to measure the similarity between the transformed feature and each prototype. The resulting soft assignment is given by

$$\mathbf{E}_{imk}^v = \frac{\left(1+\|f_m(\mathbf{z}_i^v)-\boldsymbol{\mu}_k\|^2\right)^{-1}}{\sum_{j=1}^{C}\left(1+\|f_m(\mathbf{z}_i^v)-\boldsymbol{\mu}_j\|^2\right)^{-1}}. \tag{3}$$

Here, $f_m(\cdot)$ captures expert-specific clustering preferences, while the shared prototypes enforce semantic alignment across experts and views. By aggregating $\mathbf{E}_{imk}^v$ over all samples, we obtain the expert clustering output $\mathbf{E}_m^v$.

Gating Network for Expert Selection. To adaptively combine multiple clustering experts, MoCE employs a shared gating network $\pi_\psi(\cdot)$ to generate expert-wise routing weights. For each sample i and view v, the gating weight assigned to expert m is defined as:

$$\mathbf{G}_{im}^v = \frac{\exp\left(\pi_\psi(\mathbf{z}_i^v)_m\right)}{\sum_{j=1}^{M}\exp\left(\pi_\psi(\mathbf{z}_i^v)_j\right)}. \tag{4}$$

This gating mechanism enables the model to dynamically select the most relevant experts for each input. To further improve routing sparsity and computational efficiency, a top-K routing strategy is adopted, where only the K experts with the highest routing probabilities are activated.

2.2 Dual-Level Multi-View Consistency Learning

Although MoCE enhances intra-cluster modeling by allowing multiple experts to capture diverse substructures, this increased flexibility may amplify semantic discrepancies across views in unsupervised multi-view settings. To address this issue, we propose a dual-level consistency learning framework, which enforces alignment from both the sample level and the global semantic level. The sample-level mechanism ensures balanced and consistent cluster assignments across views, while the global-level supervision provides a unified semantic anchor to prevent accumulated view bias.

At the sample level, we aim to align clustering assignments across views while avoiding degeneration of the shared clustering prototypes. To achieve this, we introduce a balanced assignment strategy combined with an alternating prediction mechanism. Specifically, for each view v, we compute a balanced assignment $\widetilde{\mathbf{Q}}^v \in \mathbb{R}^{N\times C}$ by solving the following entropy-regularized optimization problem:

$$\begin{aligned} &\widetilde{\mathbf{Q}}^v = \arg\min_{\mathbf{Q}} \left\langle \mathbf{Q}, -\log \mathbf{Q}^v \right\rangle + \varepsilon\, H(\mathbf{Q}), \\ &s.t.\ \mathbf{Q}\mathbf{1}_C = \frac{1}{N}\mathbf{1}_N,\ \mathbf{Q}^\top \mathbf{1}_N = \frac{1}{C}\mathbf{1}_C \end{aligned} \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the matrix inner product, $H(\mathbf{Q})$ is the entropy regularization term, and ε is a hyperparameter. We have provided a more detailed optimization process in supplementary materials. This formulation enforces balanced cluster utilization, ensuring that each cluster receives a comparable number of samples. The resulting distribution is smoother and more stable, facilitating reliable cross-view alignment.

Based on the balanced assignments $\{\widetilde{\mathbf{Q}}^v\}_{v=1}^V$, we adopt an alternating prediction strategy in which each sample's soft cluster assignment in one view is aligned with the average of its assignments in the other views. For each view v, a consensus target is constructed by averaging the balanced predictions from all other views:

$$\bar{\mathbf{Q}}^{-v} = \frac{1}{V-1} \sum_{\substack{v'=1 \\ v' \neq v}}^{V} \widetilde{\mathbf{Q}}^{v'}. \tag{6}$$

The alternating prediction loss then minimizes the cross-entropy between the soft cluster assignment $\mathbf{Q}^v$ and this target $\bar{\mathbf{Q}}^{-v}$:

$$\mathcal{L}_{\mathrm{s}} = \frac{1}{V} \sum_{v=1}^{V} \sum_{i=1}^{N} \sum_{k=1}^{C} \bar{q}_{ik}^{-v} (-\log q_{ik}^v). \tag{7}$$

Through this alternating alignment mechanism, clustering predictions from different views are progressively aligned, while avoiding direct pairwise matching that may propagate noisy errors.

While sample-level alignment improves local consistency, clustering decisions based solely on individual views may suffer from view bias, as the view-specific representation $\mathbf{z}^v$ captures only information from the v-th view. To provide a more discriminative and view-invariant semantic reference, we further introduce a global target distribution from the global multi-view representation.

For each sample $\mathbf{x}_i$, we form a global representation by concatenating its view-specific representations:

$$\mathbf{h}_i = \left[\mathbf{z}_i^1; \mathbf{z}_i^2; \ldots; \mathbf{z}_i^V\right] \in \mathbb{R}^D, \tag{8}$$

where $D = \sum_{v=1}^V d_v$. Let $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]^\top \in \mathbb{R}^{N \times D}$ denote the global representation. We then perform K-means clustering on $\mathbf{H}$ to obtain global cluster centroids $\{\mathbf{c}_j\}_{j=1}^C$. The soft cluster assignment of $\mathbf{h}_i$ to the j-th global cluster is computed via the Student's t-distribution:

$$s_{ij} = \frac{\left(1 + \|\mathbf{h}_i - \mathbf{c}_j\|^2\right)^{-1}}{\sum_{l=1}^{C} \left(1 + \|\mathbf{h}_i - \mathbf{c}_l\|^2\right)^{-1}}, \tag{9}$$

yielding a soft assignment matrix $\mathbf{S} \in \mathbb{R}^{N \times C}$. To emphasize high-confidence assignments, we sharpen $\mathbf{S}$ into a target distribution $\mathbf{P}$ as follows:

$$p_{ij} = \frac{s_{ij}^2 / \sum_{i=1}^N s_{ij}}{\sum_{j'=1}^{C} \left(s_{ij'}^2 / \sum_{i=1}^N s_{ij'}\right)}. \tag{10}$$

The global target distribution $\mathbf{P}$ provides a unified supervisory signal that encodes the consensus clustering structure across all views.

We enforce each view's cluster assignment $\mathbf{Q}^v$ to align with $\mathbf{P}$ via a global consensus alignment loss, defined as the sum of Kullback–Leibler (KL) divergences:

$$\mathcal{L}_{\mathrm{g}} = \sum_{v=1}^{V} D_{\mathrm{KL}}\big(\mathbf{P} \,\|\, \mathbf{Q}^v\big) = \sum_{v=1}^{V} \sum_{i=1}^{N} \sum_{k=1}^{C} p_{ik} \log \frac{p_{ik}}{q_{ik}^v}. \tag{11}$$

This loss encourages the view-specific clustering assignment $\mathbf{Q}^v$ to align with the consensus semantics encoded in $\mathbf{P}$, thereby reducing view-specific bias.

The overall clustering objective is formulated by jointly minimizing the sample-level alternating prediction loss and the global consensus alignment loss:

$$\mathcal{L}_{\mathrm{C}} = \mathcal{L}_{\mathrm{s}} + \mathcal{L}_{\mathrm{g}}. \tag{12}$$

2.3 Gating Load Balance

During training, mixture-of-experts (MoE) models often suffer from expert load imbalance: a large fraction of input instances are consistently routed to only a few experts, while the majority remain underutilized. In multi-view clustering, this issue is exacerbated by the absence of ground-truth labels, as the gating network tends to converge to degenerate solutions that over-rely on a sparse subset of experts, thereby undermining model capacity and assignment diversity.

To mitigate this, we introduce an expert load balancing regularization term. Consider a model with M experts and a dataset of N samples across V views, let $T = N \times V$ denote the total number of sample–view pairs. For each expert m, we define its routing frequency as:

$$f_m = \frac{1}{T} \sum_{i=1}^{N} \sum_{v=1}^{V} \mathbb{I}\big[m \in \mathcal{T}_K(\mathbf{g}_i^v)\big], \tag{13}$$

where $\mathbf{g}_i^v = [g_{i1}^v, \ldots, g_{iM}^v]^\top$ denotes the gating distribution for sample i in view v, and $\mathcal{T}_K(\mathbf{g}_i^v)$ returns the indices of the top-K experts with the highest routing probabilities. f_m measures how often each expert is actually activated under sparse top-K routing. In parallel, we compute the mean routing probability assigned to expert m:

$$p_m = \frac{1}{T} \sum_{i=1}^{N} \sum_{v=1}^{V} g_{im}^v. \tag{14}$$

While f_m captures hard activation statistics, p_m reflects the soft routing tendency of the gating network. Based on these two complementary statistics, we formulate the gating load balancing loss as:

$$\mathcal{L}_{\mathrm{B}} = \frac{1}{M} \sum_{m=1}^{M} f_m p_m. \tag{15}$$

This objective penalizes skewed routing patterns by assigning stronger regularization to experts that are both frequently activated and assigned high routing probability. Overused experts receive stronger corrective signals, whereas underutilized experts are penalized less aggressively.

Algorithm 1 Training Procedure for MoCE-DC

Input: Multi-view data $\{\mathbf{X}^v\}_{v=1}^V$, number of clusters C, number of experts M, hyperparameters α, β.
Output: Clustering results $\{y_i\}_{i=1}^N$.

```
Pretrain autoencoders for each view by minimizing the reconstruction loss Eq. (1).
while not converged do
    Update target distribution P using current features by Eq. (10).
    for fixed target distribution P do
        for each view v = 1 to V do
            Extract latent representation Z^v ← D_e^v(X_i^v).
            Compute expert outputs E_m^v via Eq. (3).
            Compute routing weight G_m^v via Eq. (4).
            Compute view-specific soft clustering assignment Q^v via Eq. (2).
        end for
        Compute total loss: L ← L_R + αL_C + βL_B.
        Update all parameters via backpropagation.
    end for
end while
return {y_i}_{i=1}^N.
```

2.4 Joint Optimization Objective

The overall training objective of MoCE-DC integrates three complementary components:

$$\mathcal{L} = \mathcal{L}_{\mathrm{R}} + \alpha\,\mathcal{L}_{\mathrm{C}} + \beta\,\mathcal{L}_{\mathrm{B}}, \tag{16}$$

where $\mathcal{L}_{\mathrm{R}}$ denotes the multi-view reconstruction loss, $\mathcal{L}_{\mathrm{C}}$ is the multi-view clustering loss, and $\mathcal{L}_{\mathrm{B}}$ is the expert load balancing loss. The hyperparameters $\alpha, \beta > 0$ control the relative importance of each term. When the network training is completed, the final clustering result can be obtained by

$$y_i = \arg\max_j \left(\frac{1}{V} \sum_{v=1}^{V} q_{ij}^v \right). \tag{17}$$

The training procedure is summarized in Algorithm 1.

Complexity Analysis. Let N, V, C, M, $|\mathcal{B}|$, and d_h denote the number of samples, views, clusters, experts, batch size, and latent dimension, respectively. The per-batch computational cost consists of (i) view-specific encoding at $\mathcal{O}(|\mathcal{B}| \sum_{v=1} d_v d_h)$, and (ii) MoCE routing and prototype assignment at $\mathcal{O}(|\mathcal{B}| V M C d_h)$. Since the gating, aggregation, and consistency losses are asymptotically dominated by these terms, the per-batch complexity simplifies to $\mathcal{O}\big(|\mathcal{B}|(\sum_{v=1} d_v d_h + V M C d_h)\big)$. Consequently, the per-epoch training cost scales linearly with N. The periodic K-means update for the target distribution incurs an additional $\mathcal{O}(T_k N C V d_h)$ per update cycle, which remains negligible due to its infrequent execution.

3 Experiments

In this section, we compare our method with the state-of-the-art clustering methods on benchmark datasets. The effectiveness of the method is demonstrated through extensive experimentation.

Dataset	Samples	Views	Clusters	Dimensions
BDGP	2500	2	5	1750/79
Caltech-5V	1400	5	7	40/254/1984/512/928
OutdoorScene	2688	4	8	512/432/256/48
CiteSeer	3312	2	6	3703/4732
Handwritten	2000	6	10	240/76/216/47/64/6
NUSWIDEOBJ	30000	5	31	65/226/145/74/129
CIFAR10	50000	3	10	512/2048/1024

Table 1: Overview of the multi-view datasets.

3.1 Datasets

To validate the effectiveness of MoCE-DC, we conduct experiments on seven benchmark datasets as shown in Table 1. BDGP [Cai *et al.*, 2012], Caltech-5V [Fei-Fei *et al.*, 2004], OutdoorScene [Fang *et al.*, 2023a], CiteSeer [Fang *et al.*, 2023a], Handwritten [Du *et al.*, 2025], NUSWIDEOBJ [Zhao *et al.*, 2025], and CIFAR10 [Yan *et al.*, 2025]. We further construct Caltech-4V, Caltech-3V, and Caltech-2V datasets based on Caltech-5V to evaluate the robustness of MoCE-DC.

3.2 Experimental Setup

We compare MoCE-DC with two single-view clustering algorithms and ten state-of-the-art multi-view clustering algorithms. The two single-view clustering algorithms are K-means [McQueen, 1967] and IDEC [Guo *et al.*, 2017]. The multi-view clustering methods can be broadly categorized into two groups. The first group comprises graph-based multi-view clustering methods, including CAMVC [Zhang *et al.*, 2024], DMAC [Wang *et al.*, 2025], GMVC [He *et al.*, 2025], ALPC [Chen *et al.*, 2025]. The second group includes end-to-end multi-view clustering approaches with explicit clustering layers, including SDMVC [Xu *et al.*, 2022a], MFLVC [Xu *et al.*, 2022b], CVCL [Chen *et al.*, 2023], DealMVC [Yang *et al.*, 2023], ADMC [Zhao *et al.*, 2024], SSLNMVC [Yan *et al.*, 2025].

For MoCE-DC, the autoencoder consists of an encoder with layer sizes 2000-500-500-d_h and a decoder with layer sizes d_h-500-500-2000. The low dimension d_h is set to 10. The batch size is set to 512, and training is performed for up to 10,000 epochs or until early stopping is triggered. For other compared methods, we use the public codes and default settings from the original papers to conduct experiments. All experiments are conducted on an NVIDIA GeForce RTX-3090 GPU. The code is available at `https://github.com/Systemclock/MoCE-DC`.

We apply three wildly used metrics to evaluate the final clustering performance, including clustering accuracy (ACC), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI). For all methods, we ran ten independent runs and reported the mean values.

3.3 Experimental Results

As shown in Table 2 and Table 3, MoCE-DC achieves the best or second-best clustering performance on all seven multi-view datasets, demonstrating strong stability and generalization ability. First, compared with single-view clus-

Methods	BDGP			OutdoorScene			CiteSeer			Handwritten			NUSWIDEOBJ		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
KM	52.09	34.87	18.24	44.00	33.94	23.28	32.96	12.41	7.78	68.49	65.81	55.22	11.80	10.47	3.18
IDEC	70.51	55.05	52.89	53.61	48.74	34.89	57.19	29.35	28.88	68.77	69.34	57.73	11.10	9.17	2.97
CAMVC	93.96	82.67	83.85	65.36	50.85	42.67	59.18	31.15	30.90	92.80	86.17	84.80	16.08	15.88	6.24
DMAC	93.24	87.33	84.85	55.88	46.09	35.71	37.56	16.53	12.13	82.93	74.81	68.67	12.50	12.69	4.38
GMVC	96.35	92.84	92.63	64.15	53.71	41.50	38.03	17.44	14.10	77.36	75.23	65.21	14.96	**17.36**	6.84
ALPC	88.20	76.26	71.41	36.17	26.60	17.30	44.14	19.86	18.63	62.25	57.90	43.69	14.57	8.21	3.33
SDMVC	82.84	70.79	68.06	68.79	58.37	48.22	59.63	32.76	33.25	96.90	94.54	93.54	15.01	13.97	6.60
MFLVC	97.83	94.24	95.86	59.39	52.45	47.42	42.77	22.41	32.65	77.86	79.99	77.27	19.98	14.71	6.32
CVCL	92.72	81.23	82.92	63.55	53.36	39.93	41.34	20.40	16.27	96.00	92.73	91.74	11.92	11.86	3.97
DealMVC	66.04	62.10	49.99	53.58	51.72	38.53	56.70	32.06	43.04	80.13	79.84	71.06	17.62	10.52	2.67
ADMC	90.16	74.29	77.00	71.35	59.27	50.70	56.82	26.07	24.33	90.75	83.82	80.91	14.18	7.43	2.60
SSLNMVC	98.68	96.09	96.74	62.09	53.95	40.82	29.50	11.12	7.51	83.10	81.39	74.28	20.90	13.80	**8.88**
MoCE-DC	**99.20**	**97.64**	**98.02**	**71.58**	**59.31**	**50.75**	**61.08**	**33.97**	**33.73**	**98.95**	**97.59**	**97.69**	**20.97**	14.37	7.73

Table 2: The clustering results with ACC, NMI and ARI. The best result is indicated in bold, while the second-ranked result is marked with an underline.

Methods	Caltech-5V			Caltech-4V			Caltech-3V			Caltech-2V			CIFAR10		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
KM	60.97	48.85	41.52	61.14	48.61	41.33	61.22	48.49	42.32	49.30	34.97	27.71	86.82	75.55	72.86
IDEC	60.02	52.14	44.30	62.11	53.51	45.92	59.80	49.87	43.06	47.98	36.84	28.78	75.80	64.03	57.30
CAMVC	86.29	76.27	72.35	91.36	83.99	82.45	88.29	79.89	76.57	60.57	47.49	40.92	99.41	98.31	**98.70**
DMAC	50.57	36.04	28.76	59.79	47.36	36.11	62.43	49.21	38.55	58.71	42.13	34.27	95.86	91.68	91.23
GMVC	81.94	75.56	70.43	85.09	75.83	72.39	78.49	69.64	64.61	61.27	53.84	43.31	99.24	97.88	98.34
ALPC	47.86	28.25	21.83	48.07	28.33	21.82	47.36	27.06	20.99	47.29	26.62	20.84	96.83	92.74	93.24
SDMVC	90.22	81.39	80.15	84.71	72.61	70.13	83.07	70.79	67.06	43.57	41.86	29.62	58.74	64.44	60.20
MFLVC	70.45	61.54	60.71	83.92	74.17	73.99	70.73	62.16	60.97	60.99	53.43	53.04	99.31	98.09	98.64
CVCL	87.79	79.79	79.22	78.91	65.94	66.45	70.51	60.77	60.34	53.71	43.90	44.15	99.27	97.97	98.40
DealMVC	87.19	79.20	75.72	73.71	64.79	57.33	63.44	53.21	43.89	47.34	40.59	30.19	91.92	90.57	87.32
ADMC	91.64	84.81	82.56	92.64	87.03	85.24	88.14	80.25	77.53	62.50	52.60	41.87	87.68	82.16	73.79
SSLNMVC	78.64	72.44	66.20	89.00	80.69	77.25	78.36	69.05	62.42	64.36	53.98	44.20	99.22	97.88	98.31
MoCE-DC	**92.86**	**87.23**	**85.49**	**94.29**	**88.60**	**87.96**	**92.14**	**85.90**	**84.22**	**69.00**	**56.33**	**49.71**	**99.46**	**98.55**	98.64

Table 3: The clustering results with ACC, NMI and ARI. The best result is indicated in bold, while the second-ranked result is marked with an underline.

tering methods, MoCE-DC consistently yields superior results, as it effectively exploits the complementary information across multiple views. Second, unlike methods that obtain cluster assignments through a single MLP-based clustering head, MoCE-DC explicitly models intra-cluster substructures by introducing multiple collaborative experts and adaptively derives the final cluster assignments, which enhances its capability to capture intra-cluster diversity. Finally, although graph-based methods can also reveal latent substructures within clusters to some extent, their performance heavily depends on the quality of graph construction. By contrast, MoCE-DC achieves cross-view alignment through a hierarchical consistency constraint, leading to more robust and improved clustering performance.

3.4 Ablation Study

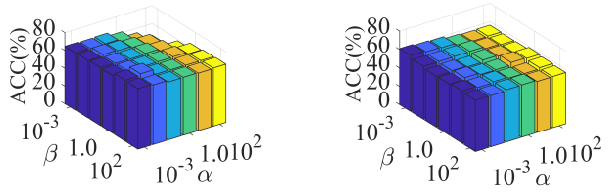
Table 4 reports the ablation study results, evaluating the contribution of each loss component. First, when only the reconstruction loss $\mathcal{L}_R$ is used, the clustering performance is extremely poor, indicating that representation learning alone is insufficient to induce meaningful clustering structures. Second, by directly supervising the formation of clustering assignments, the clustering loss $\mathcal{L}_C$ enables the model to learn more discriminative representations. Third, the gating balance regularization $\mathcal{L}_B$ can also improve clustering performance. Overall, the best performance is consistently achieved when all loss terms are jointly optimized, confirming the necessity of jointly modeling clustering supervision and consistency alignment in the proposed MoCE-DC framework.

3.5 Parameter Analysis

For the proposed MoCE-DC framework, we investigate the sensitivity of key parameters. Specifically, α and β are selected via grid search from the set $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$. The number of experts is chosen from $\{C, 10, 15, 20, 25\}$. As shown in Figure 3, MoCE-DC exhibits relatively stable clustering performance across a wide range of α and β values, indicating that the proposed framework is not overly sensitive to the precise choice of these hyperparameters. As shown in Figure 4, the cluster-

Dataset	L_R	L_C	L_B	ACC	NMI	ARI
BDGP	✓	✗	✗	39.68	21.92	10.8
	✓	✓	✗	96.92	94.75	95.33
	✓	✗	✓	37.16	20.02	8.12
	✓	✓	✓	**99.20**	**97.64**	**98.02**
Caltech-5V	✓	✗	✗	27.00	8.04	4.26
	✓	✓	✗	91.43	85.39	83.80
	✓	✗	✓	32.29	15.88	10.61
	✓	✓	✓	**92.86**	**87.23**	**85.49**
OutdoorScene	✓	✗	✗	25.78	11.78	4.29
	✓	✓	✗	67.44	56.21	46.88
	✓	✗	✓	23.07	9.34	2.75
	✓	✓	✓	**71.58**	**59.31**	**50.75**
CiteSeer	✓	✗	✗	30.43	6.29	3.94
	✓	✓	✗	58.72	30.96	31.69
	✓	✗	✓	28.71	5.31	2.92
	✓	✓	✓	**61.08**	**33.97**	**33.73**

Table 4: The clustering results of Ablation Study.



(a) OutdoorScene (b) CiteSeer

Figure 3: Clustering ACC(%) obtained under various settings of the parameters α and β.

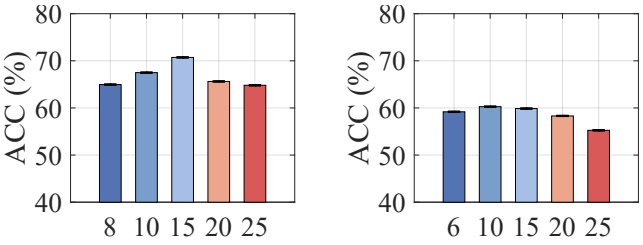
ing performance reaches its peak when the number of experts M is set to approximately 10 or 15. Increasing M beyond this range results in degraded performance, because an excessive number of experts introduces redundancy and weakens effective expert cooperation. Therefore, we set M within the range of $2C$ to $3C$ to strike a balance between model capacity and clustering stability.

3.6 Convergence Analysis

As shown in Figure 5, we visualize the clustering metrics and overall loss throughout training to validate the convergence. The total loss drops quickly and then plateaus, and the clustering accuracy correspondingly converges to a stable level. When the target distribution is updated, a temporary increase in the loss can be observed. However, the model subsequently converges to a better clustering solution, indicating that the global target distribution effectively guides the views toward learning more accurate clustering assignments.

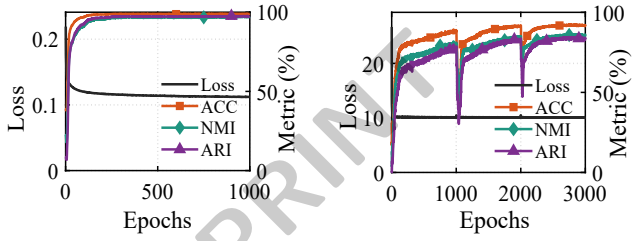
3.7 Visualization Validation

As shown in Figure 6, we perform t-SNE visualization of the learned global representations, where each sample is colored according to its final clustering result. The visualization shows clear inter-cluster boundaries and compact intra-cluster distributions on both BDGP and Caltech-5V. This demonstrates that MoCE-DC effectively enhances the discriminability of the learned representation and produces reliable clustering assignments.



(a) OutdoorScene (b) CiteSeer

Figure 4: Clustering ACC(%) obtained with different number of experts. The horizontal axis represents the number of selected experts.

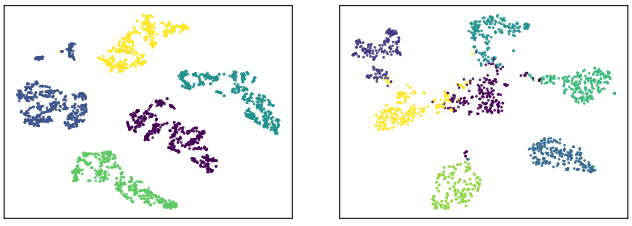


(a) BDGP (b) Caltech-5V

Figure 5: Convergence curves and clustering performance on BDGP and Caltech-5V.

4 Conclusion

This paper proposes a novel multi-view clustering framework, termed MoCE-DC, which integrates a mixture of clustering experts with a dual-level consistency constraint. MoCE-DC decomposes complex cluster structures by leveraging multiple experts to capture diverse sub-cluster patterns within each view. To ensure cross-view consistency, MoCE-DC adopts a dual-level alignment strategy, including sample-wise cross-view prediction and global consensus regularization. Extensive experimental results demonstrate the effectiveness of the proposed method.



(a) BDGP (b) Caltech-5V

Figure 6: t-SNE analysis of MoCE-DC on BDGP and Caltech-5V datasets.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2022ZD0160402), the National Natural Science Foundation of China (No. 62201471), and Aeronautical Science Fund (No. ASFC-20240001053006).

Contribution Statement

All authors made contributions to the paper's methodology and experiments. Daidai Zhu and Yang Zhao contributed equally to this work and should be considered co-first authors. Yang Zhao is the corresponding author.

References

[Bai *et al.*, 2024] Ruina Bai, Ruizhang Huang, Yanping Chen, Yongbin Qin, Yong Xu, and Qinghua Zheng. A hierarchical consensus learning model for deep multi-view document clustering. *Information Fusion*, 111:102507, 2024.

[Cai *et al.*, 2012] Xiao Cai, Hua Wang, Heng Huang, and Chris Ding. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12):i16–i24, 2012.

[Chao *et al.*, 2021] Guoqing Chao, Shiliang Sun, and Jinbo Bi. A survey on multiview clustering. *IEEE transactions on artificial intelligence*, 2(2):146–168, 2021.

[Chen *et al.*, 2023] Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16752–16761, 2023.

[Chen *et al.*, 2025] Yawei Chen, Huibing Wang, Jinjia Peng, and Yang Wang. Anchor learning with potential cluster constraints for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 15939–15947, 2025.

[Du *et al.*, 2025] Liang Du, Yukai Shi, Yan Chen, Feijiang Li, Peng Zhou, Lei Duan, and Yuhua Qian. A dual mixture-of-experts framework for multi-view k-means clustering with view balance and regional sparsity. *Information Fusion*, page 103449, 2025.

[Fang *et al.*, 2023a] Si-Guo Fang, Dong Huang, Xiao-Sha Cai, Chang-Dong Wang, Chaobo He, and Yong Tang. Efficient multi-view clustering via unified and discrete bipartite graph learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):11436–11447, 2023.

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

[Fei-Fei *et al.*, 2004] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[Feng *et al.*, 2024] Wei Feng, Guoshuai Sheng, Qianqian Wang, Quanxue Gao, Zhiqiang Tao, and Bo Dong. Partial multi-view clustering via self-supervised network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 11988–11995, 2024.

[Guo *et al.*, 2017] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Ijcai*, volume 17, pages 1753–1759, 2017.

[He *et al.*, 2025] Hongqing He, Jie Xu, Guoqiu Wen, Yazhou Ren, Na Zhao, and Xiaofeng Zhu. Graph embedded contrastive learning for multi-view clustering. In *Proceedings of the thirty-fourth international joint conference on artificial intelligence, IJCAI-25*, pages 5336–5344. International Joint Conferences on Artificial Intelligence Organization, 2025.

[Huang *et al.*, 2023] Zongmo Huang, Yazhou Ren, Xiaorong Pu, Shudong Huang, Zenglin Xu, and Lifang He. Self-supervised graph attention networks for deep weighted multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7936–7943, 2023.

[Ke *et al.*, 2021] Guanzhou Ke, Zhiyong Hong, Zhiqiang Zeng, Zeyi Liu, Yangjie Sun, and Yannan Xie. Conan: contrastive fusion networks for multi-view clustering. In *2021 IEEE International conference on big data (Big Data)*, pages 653–660. IEEE, 2021.

[Ling *et al.*, 2023] Yawen Ling, Jianpeng Chen, Yazhou Ren, Xiaorong Pu, Jie Xu, Xiaofeng Zhu, and Lifang He. Dual label-guided graph refinement for multi-view graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 8791–8798, 2023.

[McQueen, 1967] James B McQueen. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pages 281–297, 1967.

[Pan and Kang, 2021] Erlin Pan and Zhao Kang. Multi-view contrastive graph clustering. *Advances in neural information processing systems*, 34:2148–2159, 2021.

[Wang *et al.*, 2021] Yiming Wang, Dongxia Chang, Zhiqiang Fu, and Yao Zhao. Consistent multiple graph embedding for multi-view clustering. *IEEE transactions on multimedia*, 25:1008–1018, 2021.

[Wang *et al.*, 2023] Yiming Wang, Dongxia Chang, Zhiqiang Fu, and Yao Zhao. Learning a bi-directional discriminative representation for deep clustering. *Pattern Recognition*, 137:109237, 2023.

[Wang *et al.*, 2025] Bocheng Wang, Chusheng Zeng, Mulin Chen, and Xuelong Li. Towards learnable anchor for deep multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21044–21052, 2025.

[Xu *et al.*, 2021] Jie Xu, Yazhou Ren, Guofeng Li, Lili Pan, Ce Zhu, and Zenglin Xu. Deep embedded multi-view clustering with collaborative training. *Information Sciences*, 573:279–290, 2021.

[Xu *et al.*, 2022a] Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, Philip S Yu, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7470–7482, 2022.

[Xu *et al.*, 2022b] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16051–16060, 2022.

[Yan *et al.*, 2023] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19863–19872, 2023.

[Yan *et al.*, 2025] Weiqing Yan, Tingyu Yang, and Chang Tang. Self-supervised semantic soft label learning network for deep multi-view clustering. *IEEE Transactions on Multimedia*, 2025.

[Yang *et al.*, 2021] Xu Yang, Cheng Deng, Zhiyuan Dang, and Dacheng Tao. Deep multiview collaborative clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):516–526, 2021.

[Yang *et al.*, 2023] Xihong Yang, Jin Jiaqi, Siwei Wang, Ke Liang, Yue Liu, Yi Wen, Suyuan Liu, Sihang Zhou, Xinwang Liu, and En Zhu. Dealmvc: Dual contrastive calibration for multi-view clustering. In *Proceedings of the 31st ACM international conference on multimedia*, pages 337–346, 2023.

[Zhang *et al.*, 2019] Changqing Zhang, Yeqing Liu, and Huazhu Fu. Ae2-nets: Autoencoder in autoencoder networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2577–2585, 2019.

[Zhang *et al.*, 2024] Chao Zhang, Xiuyi Jia, Zechao Li, Chunlin Chen, and Huaxiong Li. Learning cluster-wise anchors for multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16696–16704, 2024.

[Zhao *et al.*, 2024] Helin Zhao, Wei Chen, and Peng Zhou. Active deep multi-view clustering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 5554–5562, 2024.

[Zhao *et al.*, 2025] Yang Zhao, Daidai Zhu, Aihong Yuan, and Xuelong Li. Consensus multi-view spectral clustering network with unified similarity. *Neural Networks*, 190:107647, 2025.

[Zhou *et al.*, 2024] Lihua Zhou, Guowang Du, Kevin Lue, Lizheng Wang, and Jingwei Du. A survey and an empirical evaluation of multi-view clustering approaches. *ACM Computing Surveys*, 56(7):1–38, 2024.