Mixture of Experts as Representation Learner for Deep Multi-View Clustering

Yunhe Zhang[1,*], Jinyu Cai[2,*,†], Zhihao Wu[3], Pengyang Wang[1], See-Kiong Ng[2]

[1]Department of Computer and Information Science, SKL-IOTSC, University of Macau, China
[2]Institute of Data Science, National University of Singapore, Singapore
[3]College of Computer Science and Technology, Zhejiang University, China
{zhangyhannie, jinyucai1995, zhihaowu1999}@gmail.com, pywang@um.edu.mo, seekiong@nus.edu.sg

Abstract

Multi-view clustering (MVC) aims to integrate information from diverse data sources to facilitate the clustering process, which has achieved considerable success in various real-world applications. However, previous MVC methods typically employ one of two strategies: (1) designing separate feature extraction pipelines for each view, which restricts their ability to fully exploit collaborative potential; or (2) employing a single shared representation module, which hinders the capture of diverse, view-specific representations. To tackle these challenges, we introduce Deep Multi-View Clustering via Collaborative Experts (DMVC-CE), a novel MVC approach that employs the Mixture of Experts (MoE) framework. DMVC-CE incorporates a gating network that dynamically selects multiple experts for handling each data sample, capturing diverse and complementary information from different views. Additionally, to ensure balanced expert utilization and maintain their diversity, we introduce an equilibrium loss and a multi-expert distinctiveness enhancer. The equilibrium loss prevents excessive reliance on specific experts, while the distinctiveness enhancer encourages each expert to specialize in different aspects of the data, thereby promoting diversity in learned representations. Comprehensive experiments on various multi-view benchmark datasets demonstrate the superiority of DMVC-CE compared to state-of-the-art MVC baselines.

Introduction

Clustering (Xu and Wunsch 2005; Cai et al. 2022a; Wu et al. 2023; Liu et al. 2023a; Cai et al. 2024c; Ren et al. 2024) is an essential research issue in unsupervised machine learning, which has numerous applications in real-world scenarios, *e.g.*, drug discovery (Sadybekov and Katritch 2023) and social network analysis (Li et al. 2017; Cai et al. 2024b). With the rapid growth of multimedia technology, the data can be expressed via multiple perspectives, *e.g.*, text, image, graph, etc, which is so-called multi-view data. Compared to single-view data, multi-view data offer richer information for representing an object, which provides the potential to further improve clustering performance.

Multi-view clustering (MVC) (Hu et al. 2023; Yang et al. 2023; Huang, Yang, and Cai 2024) aims to integrate information from multiple views to achieve more efficient clustering than handling each view independently. Early MVC methods typically rely on classical machine learning techniques, such as multiple kernel learning (Tzortzis and Likas 2012; Houthuys, Langone, and Suykens 2018; Wu, Zhang, and Fan 2024), matrix factorization (Liu et al. 2013; Li et al. 2024), graph methods (Huang et al. 2019; Liu et al. 2024; Wen et al. 2024), and subspace learning (Sun et al. 2021; Cai et al. 2024a), to fuse features from various views. However, these methods generally face limitations in scalability and handling high-dimensional data. Fortunately, the emergence of deep learning has significantly advanced clustering methods (Qi et al. 2022; Cai et al. 2022b; Liu et al. 2023b). Deep learning-based MVC methods (Li et al. 2019; Xu et al. 2022; Fang et al. 2023; Pu et al. 2024) leverage the powerful representation learning capability of deep neural networks to process multi-view data with high-dimensionality and complex structures. These methods excel in extracting representative features and uncovering the intricate correlations among different views, facilitating a more effective information integration of multi-view data. Moreover, traditional MVC methods have also improved from their deep variants (Chen et al. 2022; Li et al. 2023a; Chen et al. 2024).

While existing MVC methods, such as (Qi et al. 2024; Zhang, Liu, and Fu 2019), predominantly emphasize developing powerful feature fusion mechanisms across multiple views, one of the critical oversights is the conventional separated design of backbone networks in existing approaches. Typically, the feature extraction process for each view operates independently, with integration occurring only at the subsequent feature fusion stage. This can significantly limit the collaborative potential between multiple views during the early feature learning phase, resulting in underutilized cross-view correlations and sub-optimal clustering performance. While another strategy (Chen et al. 2022; Wan et al. 2024, 2023) attempts to adopt a shared representation module for multi-view data, it would fail in capturing the diversity of view-specific representations. Consequently, it unveils a compelling research challenge in MVC: *How can we facilitate collaboration across multiple backbone networks while preserving the uniqueness of view-specific representations to achieve effective multi-view clustering?*

*These authors contributed equally.
†Corresponding author.

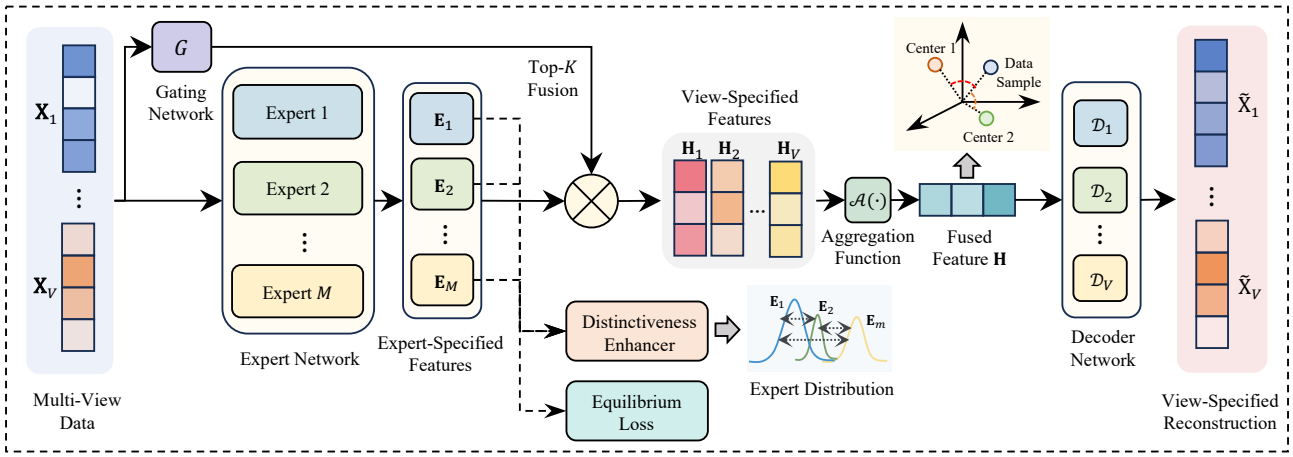


Figure 1: The architecture of the proposed DMVC-CE method. The MoE representation learner contains multiple experts and a gating network dynamically selects appropriate experts for representation learning. Besides, an equilibrium loss and a distinctiveness enhancer are introduced to ensure balanced utilization and diversity of experts. $\mathcal{A}(\cdot)$ denotes the aggregation function to integrate features among V views. A decoder network is incorporated for view-specified data reconstruction.

To address this challenge, we propose a simple yet effective MVC method called Deep Multi-View Clustering via Collaborative Experts (DMVC-CE) in this paper. Unlike the separated design in the feature extraction pipeline of existing MVC frameworks, we argue that they could be unified and encourage collaboration. We take advantage of the mixture of experts (MoE) and introduce multiple expert networks to fuse features of multi-view data. The network architecture of DMVC-CE is shown in Figure 1. Specifically, the multi-view data flow not only into multiple expert networks but also into a designed gating network, where the gating network dynamically selects experts to participate in the representation learning of each instance, thereby capturing diverse and complementary information from different views. Compared to the aforementioned strategies, the unified MOE representation learner assigns different combinations of experts to different views, enabling the optimization of these views to explicitly have a collaborative impact on the overlapping selected expert networks.

However, there are two potential risks during training: 1) certain experts may be excessively relied on; 2) different experts may converge to the same functions. To mitigate these issues, we introduce an equilibrium loss to ensure balanced expert participation. Besides that, we propose a multi-expert distinctiveness enhancer to maintain expert diversity by maximizing the mutual information between expert distributions. The main contributions of this paper are:

- We propose DMVC-CE to bridge a pivotal gap in existing MVC frameworks, *i.e.*, the limited collaborative potential between multiple feature extraction pipelines. DMVC-CE leverages the MoE architecture to facilitate collaborative processing across different views.
- We introduce an equilibrium loss for balancing the utilization of experts, which prevents excessive reliance on specific experts.
- We propose a distinctiveness enhancer to encourage expert diversity, thereby facilitating the learning of diverse information from multi-view data.
- We empirically validate the superiority of DMVC-CE over state-of-the-art MVC baselines through rigorous evaluation of various benchmark datasets.

Related Works

Multi-view Clustering

Multi-view clustering (MVC) (Fu et al. 2022; Chen et al. 2023b; Yang et al. 2023; Zhang et al. 2024; Zhou et al. 2024) aims to leverage complementary information from diverse data sources to enhance clustering performance. Traditional MVC methods mainly focus on graph-based learning, matrix factorization, subspace learning, etc. For instance, Li et al. effectively leveraged anchor graphs in multi-view clustering. Non-negative matrix factorization (Yang et al. 2020) is used to streamline large-scale multi-view clustering. Chen et al. sought for a consensus subspace to capture complementarity across views. In recent years, the rise of deep learning has further advanced MVC. Some studies encode self-expression by utilizing deep neural networks so as to implement multi-view subspace clustering (Wang et al. 2020; Zhu et al. 2024). Meanwhile, deep matrix factorization and deep multiple kernel models are also developed (Zhao, Ding, and Fu 2017; Huang et al. 2022; Li et al. 2023a). These methods significantly gain better scalability and learning capacity over traditional counterparts. Nevertheless, existing MVC methods typically employ separate feature extraction pipelines for each view, which prevents mutual collaboration between different views and leads to sub-optimal performance.

Mixture of Experts

The mixture of experts (MoE) (Yuksel, Wilson, and Gader 2012) is a machine learning paradigm designed to enhance

learning efficiency and model performance by distributing multi-tasks among multiple specialized sub-models, known as experts. Each expert is trained to focus on different aspects of the input space, while a gating network dynamically determines the contribution of each expert based on the input data. MoE has demonstrated its effectiveness across various domains. For example, in natural language processing, the switch transformer (Fedus, Zoph, and Shazeer 2022) leverages MoE to enhance the efficiency of the transformer by activating only a subset of experts for each input token. In computer vision, MoE has been utilized to improve vision tasks such as image classification (He et al. 2021) and video recognition (Li and Xu 2023) by allowing different experts to focus on different features or regions of an image.

Despite these successes, the potential of MoE in improving multi-view clustering remains under-explored. MoE offers a promising and scalable solution for multi-view clustering with the potential to effectively learn and integrate complementary information from multiple views. This paper aims to bridge this research gap by leveraging the MoE paradigm to develop a simple yet effective MVC method that fully exploits the collaborative potential of multi-view data.

Methodology

Problem Formulation

Let $\{\mathbf{X}^{(v)} = [\mathbf{x}_1^{(v)}, \ldots, \mathbf{x}_N^{(v)}]\}_{v=1}^V$ denote a multi-view dataset, where N and V are the number of data samples and views, respectively. Specifically, $\mathbf{X}^{(v)} \in \mathbb{R}^{N \times d_v}$ represents the data of v-th view, with d_v being the dimensionality of feature space in this view. The primary objective of multi-view clustering is to exploit the comprehensive information available across V views of data to learn more discriminative features than single-view data, thereby facilitating a more accurate partition of the N data samples into C clusters. Existing MVC methods typically utilize separated feature extraction pipelines for each view, followed by feature fusion and clustering processes.

Different from existing MVC methods, we propose a simple yet effective MVC framework called deep multi-view clustering via collaborative experts (DMVC-CE) in this paper, which introduces a collaborative MoE architecture as the representation learner to facilitate multi-view clustering. Our method encourages interaction and information sharing across views, thereby capturing the diverse and complementary features of the multi-view data. Formally, the overall objective function can be expressed as:

$$\mathcal{L} = \sum_{v=1}^{V} \mathcal{L}_{\text{recon}}^{(v)} + \mathcal{L}_{\text{expert}}, \tag{1}$$

where $\mathcal{L}_{\text{recon}}^{(v)}$ denotes the representation learning loss for the v-th view to ensure the learned features are informative, and $\mathcal{L}_{\text{expert}}$ is the loss function for the MoE representation learner. We will describe each component of DMVC-CE in the following sections.

Collaborative Multi-View Mixture of Experts

Existing MVC approaches follow a similar representation learning framework, *i.e.*, employing mutually independent representation learners for data from different views. We argue that this independent learning paradigm may fail to fully capture the complementary information across views, as the cross-view representation learners have no explicit interactions, leading to sub-optimal clustering performance. To address this limitation, we introduce MoE as a unified representation learner for multi-view clustering, which encourages information integration from different perspectives. Specifically, the proposed MoE representation learner consists of an expert network and a gating network, and the learning process can be formalized as follows:

$$\mathbf{H}_v = \sum_{m=1}^{M} \pi_m(\mathbf{X}^{(v)}; \boldsymbol{\theta}) \odot \mathcal{E}_m(\mathbf{X}^{(v)}; \boldsymbol{\phi}_m), \tag{2}$$

where $\odot$ denotes element-wise multiplication, $\mathbf{X}^{(v)}$ represents the data from the v-th view, $\mathbf{H}_v \in \mathbb{R}^{N \times d_h}$ is the learned feature, $\mathcal{E}_m(\cdot)$ represents the learning function of m-th expert, parameterized by $\boldsymbol{\phi}_m$, and $\pi_m(\cdot) \in \mathbb{R}^{N \times 1}$ denotes the gating weight of m-th expert parameterized by $\boldsymbol{\theta}$. Particularly, the gating network is crucial for ensuring that the most relevant experts contribute to form the fused representation. It dynamically selects the appropriate experts for each input data based on their specific characteristics. Formally, given an input $\mathbf{x}$, the gating network output for the m-th expert is defined by:

$$\pi_m(\mathbf{x}, \boldsymbol{\theta}) = \frac{\exp(\mathcal{G}_m(\mathbf{x}; \boldsymbol{\theta}))}{\sum_{i=1}^{M} \exp(\mathcal{G}_i(\mathbf{x}; \boldsymbol{\theta}))}, \forall m \in M, \tag{3}$$

where $\mathcal{G}_m(\cdot)$ denotes the gating function that is parameterized by a single-layer fully connected network. We follow (Shazeer et al. 2017) to utilize the top-K expert selection in the framework, where only the experts with top-K gating scores are selected to contribute to the representation learning of the corresponding input.

We then introduce an aggregation function $\mathcal{A}(\cdot)$ to obtain the fused representation $\mathbf{H}$ for v views, as formalized by:

$$\mathbf{H} = \mathcal{A}(\{\mathbf{H}_v\}_{v=1}^V), \tag{4}$$

where $\{\mathbf{H}_v\}_{v=1}^V$ denotes the feature set learned from V views. The aggregation function $\mathcal{A}(\cdot)$ is implemented with a concatenation strategy. Note that an essential issue in the MoE representation learner is to ensure that the learned fused features preserve the intrinsic information of the multi-view data. To this end, we introduce a decoder network $\mathcal{D}(\cdot)$ to reconstruct the original data from the fused feature $\mathbf{H}$. We aim to encourage the model to learn the intrinsic features of the input data and avoid distortion in the latent space. The reconstruction process can be defined by:

$$\tilde{\mathbf{X}}^{(v)} = \mathcal{D}_v(\mathbf{H}; \boldsymbol{\psi}_m), v = \{1, \ldots, V\}, \tag{5}$$

where $\tilde{\mathbf{X}}^{(v)} \in \mathbb{R}^{N \times d_v}$ and $D_v(\cdot)$ denote the reconstructed data and decoder network (parameterized by $\boldsymbol{\psi}_m$) for v-th view, respectively. Subsequently, we can define the following objective function $\mathcal{L}_{\text{recon}}$:

$$\mathcal{L}_{\text{recon}}^{(v)} = \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}_i^{(v)} - \tilde{\mathbf{x}}_i^{(v)}||^2, \tag{6}$$

where $\mathbf{x}_i^{(v)}$ and $\tilde{\mathbf{x}}_i^{(v)}$ indicate the i-th samples of input data $\mathbf{X}^{(v)}$ and reconstructed data $\tilde{\mathbf{X}}^{(v)}$. The reconstruction loss $\mathcal{L}_{\text{recon}}^{(v)}$ is calculated as the Frobenius norm of the difference between the input and reconstructed data, averaged over the numbers of samples.

Nevertheless, there still remain two significant challenges that potentially affect the effectiveness of the proposed MoE representation learner:

1. The model may rely excessively on certain experts while leaving others under-utilized, resulting in redundancy and inefficient exploitation of the model capabilities.
2. Multiple experts may converge to similar functions, which reduces the diversity of experts and results in suboptimal clustering performance.

Equilibrium for the Expert Selection

The first challenge requires ensuring the equilibrium in the expert selection, so that all experts can adequately participate in representation learning. To tackle this challenge, we introduce an equilibrium loss to penalize imbalances in expert utilization, thus promoting a more efficient and fair allocation of experts to each learning task.

We define the expert usage density as ρ to represent the average frequency that each expert is selected in the representation learning process:

$$\rho_i = \sum_{v=1}^{V} \sum_{j=1}^{N} \mathcal{M}_{ji}^{(v)}, \tag{7}$$

where ρ_i denotes the usage density of expert $\mathcal{E}_i$, and $\mathcal{M} \in \mathbb{R}^{V \times N \times M}$ is a one-hot mask matrix indicating which experts have been selected for handling each sample. Specifically, $\mathcal{M}$ is defined by:

$$\mathcal{M}_{ji}^{(v)} = \begin{cases} 1, & \text{if } \mathcal{E}_i \text{ is selected for } \mathbf{x}_j^{(v)}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

To further quantify the equilibrium in expert selection, we calculate the density proxy $\hat{\rho}$, which represents the mean selection probability for each expert:

$$\hat{\rho}_i = \sum_{v=1}^{V} \sum_{j=1}^{N} \pi_i(\mathbf{x}_j^{(v)}; \boldsymbol{\theta}). \tag{9}$$

As denoted in Eq. (3), $\pi_i(\mathbf{x}_j^{(v)}; \boldsymbol{\theta})$ represents the selection probability of expert $\mathcal{E}_i$ for input $\mathbf{x}_j^{(v)}$ in the gating network. Then, the equilibrium loss ℓ_{e} used to penalize significant deviations among the utilization of experts is defined by:

$$\ell_{\text{e}} = \frac{1}{M} \sum_{i=1}^{M} (\rho_i \cdot \hat{\rho}_i). \tag{10}$$

Note that a large discrepancy between the utilization of different experts (*e.g.*, some experts are consistently chosen over others) will result in a higher ℓ_{e}, indicating an imbalance. Therefore, minimizing Eq. (10) encourages a more balanced utilization of all available experts.

Algorithm 1: Training procedure of DMVC-CE.

Input: Multi-view dataset $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$, number of views V, number of experts M, number of selected experts K, number of clusters C.
Output: The cluster labels.

```
Initialize the network parameters.
while not convergence do
    Extract the fused feature H via Eqs. (2), (3) and top-K expert selection;
    Obtain reconstruction data X̃^(v) for each view via Eq. (5);
    Calculate the reconstruction loss L_recon via Eq. (6);
    Calculate the equilibrium loss ℓ_e via Eq. (10);
    Calculate the distinctiveness loss R²_H({E_m}_{m=1}^M) via Eq. (12);
    Back propagation and update network parameters, including θ, φ, and ψ.
end while
Perform k-means to obtain the clustering results.
return The cluster labels.
```

Multi-Expert Distinctiveness Enhancer

The second challenge highlights the risk of multiple experts converging to similar functions, which may potentially diminish the overall effectiveness of the model. This emphasizes the importance of maintaining expert diversity. To address this challenge, we propose a multi-expert distinctiveness enhancer, which encourages each expert to specialize in different aspects of data, thereby enhancing the capability of the expert network to learn diverse information from multi-view data.

We achieve this by introducing a pair-wise kernel regularizer to promote diversity among experts. Specifically, given two distributions of the latent representation learned by expert $\mathcal{E}_i$ and $\mathcal{E}_j$, denoted as $\mathbb{P}_{\mathbf{E}_i}$ and $\mathbb{P}_{\mathbf{E}_j}$, the regularizer is defined as:

$$\mathcal{R}_{\mathcal{H}}^2(\mathbb{P}_{\mathbf{E}_i}, \mathbb{P}_{\mathbf{E}_j}) = -\mathbb{E}_{\mathbf{e} \sim \mathbb{P}_{\mathbf{E}_i}, \hat{\mathbf{e}} \sim \mathbb{P}_{\mathbf{E}_j}}[\kappa(\mathbf{e}, \hat{\mathbf{e}})], \tag{11}$$

where $\mathbf{e}$ and $\hat{\mathbf{e}}$ are sampled from the distribution $\mathbb{P}_{\mathbf{E}_i}$ and $\mathbb{P}_{\mathbf{E}_j}$. $\mathcal{H}$ denotes the Reproducing Kernel Hilbert Space (RKHS), and $\kappa(\mathbf{x}, \mathbf{y}) = \exp(\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2})$ represents the Gaussian kernel function measuring the similarity between $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{H}$. The regularization term aims to minimize the mutual information between the two distributions, thereby promoting distinctiveness among experts. For all M experts with $\mathbb{P}_{\mathcal{E}} = \{\mathbb{P}_{\mathcal{E}_m}\}_{m=1}^{M}$, the loss function of distinctiveness enhancer is defined as:

$$\mathcal{R}_{\mathcal{H}}^2(\{\mathcal{E}_m\}_{m=1}^{M}) = \frac{2}{M(M-1)} \sum_{m<m'} \mathcal{R}_{\mathcal{H}}^2(\mathbb{P}_{\mathbf{E}_m}, \mathbb{P}_{\mathbf{E}_{m'}}). \tag{12}$$

By optimizing this loss function, the multi-expert distinctiveness enhancer ensures that each expert learns to capture unique aspects of the multi-view data, encouraging distinctiveness between the experts.

Method	Metric	ALOI	Caltech101-all	NUS-WIDE	HW	BDGP
k-means	ACC	49.62±1.57	13.64±0.33	31.64±0.85	60.26±4.91	49.13±7.74
	NMI	49.55±0.92	30.76±0.14	17.22±0.49	57.23±4.70	39.32±8.49
	ARI	35.50±1.62	8.31±0.44	8.70±0.39	47.14±4.79	20.28±8.76
MLAN (Nie, Cai, and Li 2017)	ACC	58.94±5.18	19.47±0.56	34.72±3.17	80.45±0.00	34.87±14.00
	NMI	59.37±4.31	25.87±1.56	22.84±1.96	82.78±0.00	16.05±14.56
	ARI	34.54±5.55	-0.39±0.12	13.76±3.50	74.38±0.00	12.18±11.27
MMGC (Tan et al. 2023)	ACC	81.62±0.05	—	16.37±0.54	22.73±2.07	44.72±0.13
	NMI	77.53±0.05	—	3.79±0.72	17.15±2.04	23.28±0.00
	ARI	70.56±0.09	—	21.06±0.13	2.81±0.94	17.72±0.00
MCGC (Zhan et al. 2019)	ACC	55.51±0.00	26.63±0.00	21.56±0.00	66.05±0.00	49.80±0.00
	NMI	55.41±0.00	30.45±0.00	11.79±0.00	69.38±0.00	27.41±0.00
	ARI	35.42±0.00	5.57±0.00	3.49±0.00	60.69±0.00	24.43±0.00
CGL (Li et al. 2021)	ACC	41.24±7.26	14.91±1.22	39.88±2.50	37.42±3.05	29.55±5.22
	NMI	45.55±6.01	34.06±0.22	22.72±2.62	47.66±1.35	20.44±3.84
	ARI	9.45±3.52	-0.81±0.11	17.22±2.15	10.83±1.15	2.95±0.66
CVCL (Chen et al. 2023a)	ACC	83.43±2.41	18.33±0.37	37.53±1.79	83.45±2.31	75.92±0.00
	NMI	80.49±1.33	37.82±0.35	22.47±0.54	80.36±2.05	68.26±0.00
	ARI	72.57±2.47	14.63±0.13	16.29±1.24	73.77±3.98	64.62±0.00
RCAGL (Liu et al. 2024)	ACC	55.24±0.00	36.65±0.00	39.81±0.00	62.25±0.00	51.80±0.00
	NMI	69.67±0.00	48.44±0.00	23.18±0.00	65.34±0.00	34.52 ±0.00
	ARI	50.15±0.00	**25.05±0.00**	27.58±0.00	59.07±0.00	43.25±0.00
SCM (Luo et al. 2024)	ACC	49.21±2.16	20.95±0.79	35.56±1.25	76.50±0.98	75.00±3.29
	NMI	50.43±3.01	38.56±1.38	20.54±0.58	69.23±0.75	57.32±1.45
	ARI	31.97±4.17	17.52±0.91	15.90±1.17	62.39±1.30	53.93±2.92
OMVCDR (Wan et al. 2024)	ACC	46.80±0.00	14.22±0.00	24.62±0.00	60.15±0.00	48.32±0.00
	NMI	52.07±0.00	32.90±0.00	35.81±0.00	58.47±0.00	33.19±0.00
	ARI	34.61±0.00	9.92±0.00	15.32±0.00	46.38±0.00	19.89±0.00
DMVC-CE	ACC	**91.54±0.59**	**42.54±1.99**	**54.08±1.03**	**98.75±1.01**	**80.82±0.30**
	NMI	**88.41±0.50**	**39.77±2.09**	**50.37±1.14**	**98.00±1.32**	**62.72±0.64**
	ARI	**83.81±0.55**	25.00±0.92	**36.72±1.43**	**97.40±2.02**	**58.23±0.88**

Table 1: Average ACCs, NMIs, and ARIs (in %) with standard deviation (10 trials). The best results are marked in **bold**.

Training Strategy

Based on the proposed modules outlined in the previous subsections, we formalize the overall objective function of DMVC-CE by extending Eq. (1) as follows:

$$\begin{aligned}\mathcal{L} &= \sum_{v=1}^{V} \mathcal{L}_{\text{recon}}^{(v)} + \mathcal{L}_{\text{expert}} \\ &= \sum_{v=1}^{V} \mathcal{L}_{\text{recon}}^{(v)} + \lambda \ell_{\text{e}} + \gamma \mathcal{R}_{\mathcal{H}}^{2}(\{\mathcal{E}_m\}_{m=1}^{M}),\end{aligned} \tag{13}$$

where λ and γ are two trade-off parameters to control the contributions of the equilibrium loss and distinctiveness loss. Here we summarize the effect of each loss function:

- $\mathcal{L}_{\text{recon}}^{(v)}$ represents the reconstruction loss for the v-th view, ensuring the preservation of view-specific information.
- ℓ_{e} is the equilibrium loss that balances the utilization of experts, preventing over-reliance on a subset of experts.
- $\mathcal{R}_{\mathcal{H}}^{2}(\{\mathcal{E}_m\}_{m=1}^{M})$ is the distinctiveness loss to encourage distinctiveness among experts, thereby learning diverse information from multi-view data.

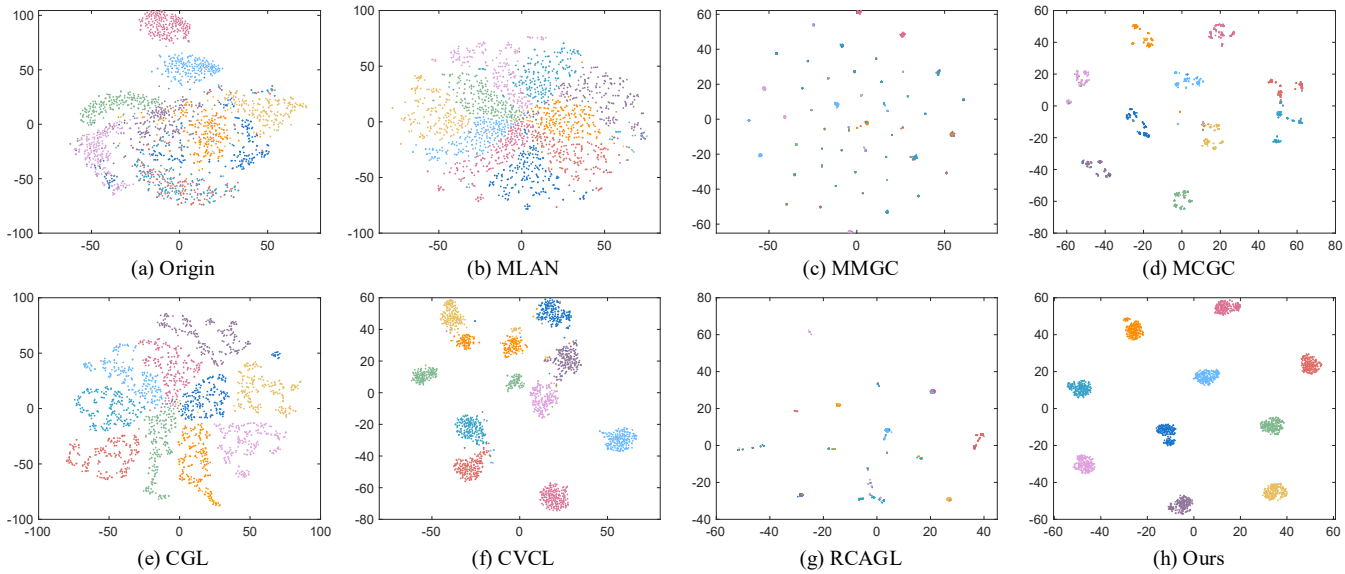
All components of DMVC-CE are trained jointly within a unified framework. After training, k-means clustering is utilized to evaluate the clustering results of the learned representations. To provide a clear understanding of the training process, the detailed training steps of DMVC-CE are presented in Algorithm 1.

Experiment

Experiment Setup

Datasets. We select five popular datasets in our experiment, including: (1) ALOI, (2) Caltech101-all, (3) NUS-WIDE, (4) HW, and (5) BDGP. Table 2 briefly summarizes the pivotal information of these datasets.

Baseline Methods. We compare the proposed DMVC-CE with (1) traditional k-means, as well as several state-of-the-



(a) Origin (b) MLAN (c) MMGC (d) MCGC

(e) CGL (f) CVCL (g) RCAGL (h) Ours

Figure 2: The t-SNE visualization results of DMVC-CE and several baseline methods on the HW dataset.

Datasets	# Total Views	# Data Samples	# Feature Dimensions	# Categories
ALOI	4	1,079	64/64/77/13	10
Caltech101-all	6	9,144	48/40/254/1,984/512/928	102
NUS-WIDE	6	1,600	64/144/73/128/225/500	8
HW	6	2,000	153/596/301/481/157/27	10
BDGP	2	2,500	1,750/79	5

Table 2: Brief illustration of the benchmark datasets.

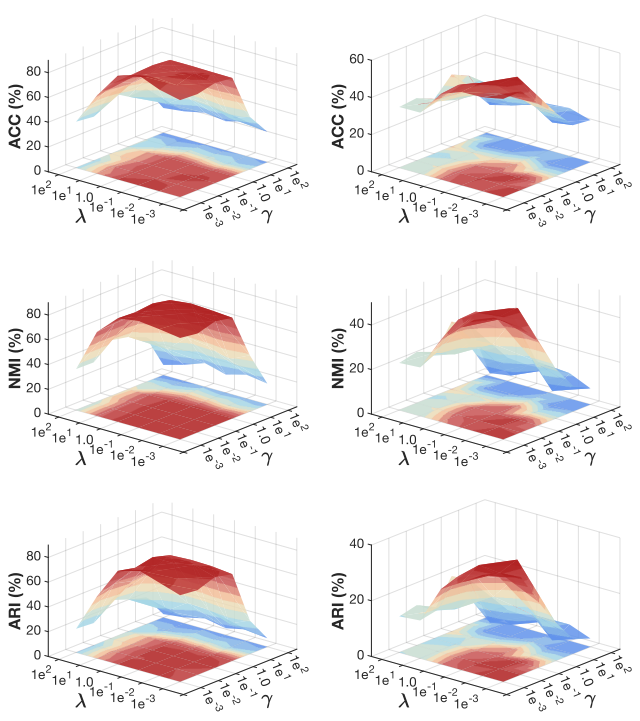
art MVC methods including: (2) MLAN (Nie, Cai, and Li 2017), (3) MMGC (Tan et al. 2023), (4) MCGC (Zhan et al. 2019), (5) CGL (Li et al. 2021), (6) CVCL (Chen et al. 2023a), (7) RCAGL (Liu et al. 2024), (8) SCM (Luo et al. 2024), (9) OMVCDR (Wan et al. 2024). All the baseline methods are performed under the default settings in the original papers. Note that the results of k-means are based on the average performance across all views.

Implementation Details. We fixed the total number of experts $M = 10$ and the selected expert number $K = 3$ in the MoE representation learner. Each expert uses a neuron setting of $2000 - 500 - 500 - d_h$, where the latent dimension d_h is fixed as 10. The network architecture of the decoder follows the same configuration as the expert. Besides, two hyper-parameters λ and γ vary in $\{0.001, 0.1, ..., 100\}$ to achieve optimal performance, and the batch size and learning rate are set to 100 and 0.005, respectively. All experiments in this paper are run on the NVIDIA Tesla A100 GPU and AMD EPYC 7532 CPU.

Evaluation Metrics. We utilize three popular evaluation metrics, including the Clustering Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI), to evaluate the clustering performance. Note that we run each method 10 times to report the clustering performance with their mean values and standard deviations.

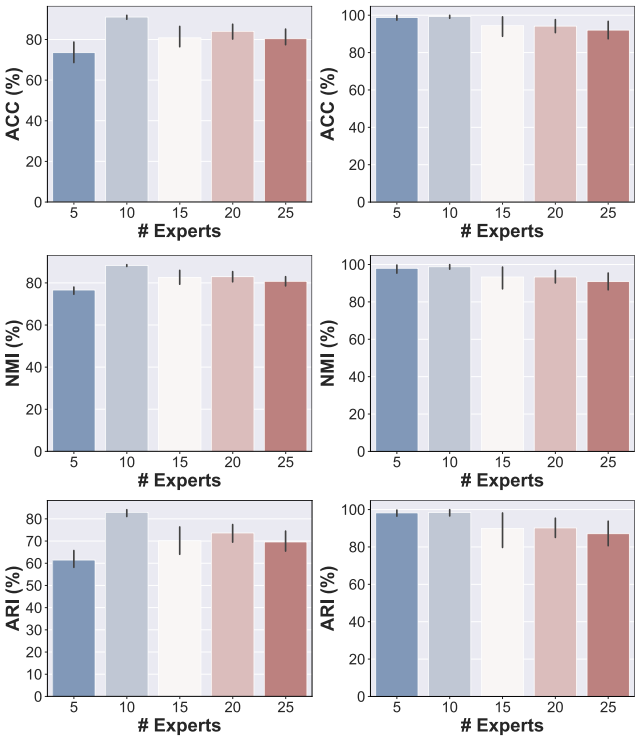
Comparison with State-of-the-art Baselines

We conduct an extensive comparison of DMVC-CE against both traditional k-means clustering and several state-of-the-art MVC methods. The experimental results, summarized in Table 1, yield the following key insights. Firstly, DMVC-CE exhibits a substantial performance improvement over the traditional k-means. This enhancement is primarily due to the capability of DMVC-CE to harness the complementary information available within multi-view data, which cannot be achieved by k-means due to its constraint of operating only on a single view. Secondly, DMVC-CE significantly outperforms other state-of-the-art MVC methods. For example, on NUS-WIDE, DMVC-CE surpasses the GCL method by a significant margin of 14.20% (ACC), 27.65% (NMI), and 19.50% (ARI), respectively. This remarkable improvement highlights the efficacy of the proposed MoE learner in representation learning. Compared to other MVC methods that employ independent feature extractors for each view, the dynamic collaboration among experts in DMVC-CE facilitates a more comprehensive integration of rich and diverse information from different views. Lastly, DMVC-CE consistently achieves superior clustering performance across all three evaluation metrics on various datasets, which demonstrates



(a) ALOI (b) NUS-WIDE

Figure 3: Performance under different values of λ and γ.



(a) ALOI (b) HW

Figure 4: Performance under different expert numbers.

its robustness and generalizability. The integration of the MoE paradigm in the DMVC-CE framework not only enhances clustering performance but also adapts well to multi-view data from different domains.

Latent Embedding Visualization

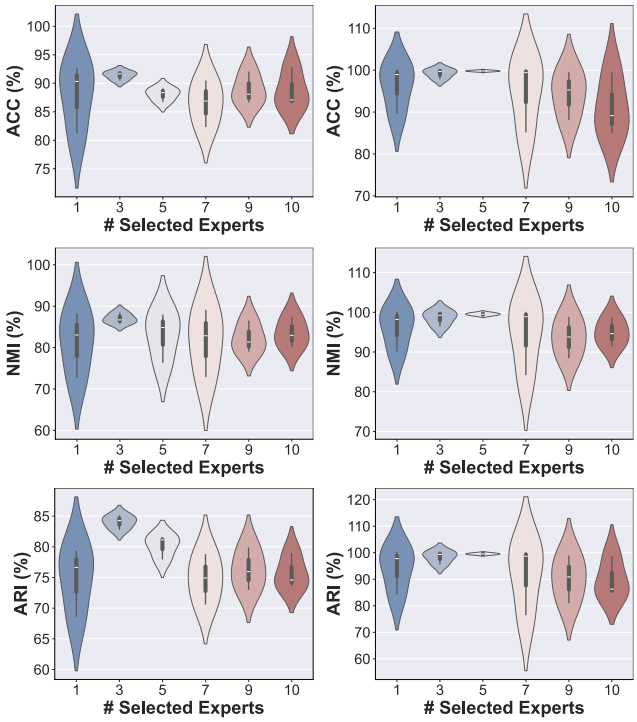
To facilitate an intuitive comparison between DMVC-CE and baseline methods, we employ t-SNE to visualize the learned latent embeddings. The visualization results, presented in Figure 2, reveal that the cluster structure of DMVC-CE is more distinct and well-separated compared to methods such as GCL or CVCL, which tend to display overlapping or poorly separated clusters. This observation indicates that DMVC-CE effectively captures the complementary information inherent in multi-view data, leading to more discriminative latent embeddings. Furthermore, DMVC-CE demonstrates tighter intra-cluster cohesion and clearer inter-cluster boundaries than the less-defined clusters observed by methods such as MMGC, MCGC, and RCAGL. This highlights that DMVC-CE more effectively integrates diverse information across multiple views.

Parameter Analysis

Impact of Hyper-Parameters λ and γ. We conduct a sensitivity analysis to evaluate the impact of the hyper-parameters λ and γ on the clustering performance. Figure 3 shows the trends of all metrics (ACC, NMI, ARI) on ALOI and NUS-WIDE datasets, with the λ and γ varying within the range of $[1e^{-3}, 1e^{2}]$. The experimental results yield several key insights. Firstly, optimal performance is observed when λ and γ fall within the range of $[1e^{-2}, 1e^{-1}]$. This finding reveals the effectiveness of both loss terms in guiding the model to learn discriminative and diverse features for multi-view data. However, it is also evident that excessively high values of λ and γ (*e.g.,* greater than $1e^{1}$) negatively impact the clustering performance. This is potentially due to an overemphasis on the corresponding loss terms during training, which can lead to imbalanced feature learning. Nevertheless, the clustering performance remains stable across a broad range of λ and γ values, which fully demonstrates the robustness of DMVC-CE.

Impact of the Total Number of Experts. To evaluate the impact of the total expert numbers M on clustering performance, we experiment by varying the total number of experts within the range of $[5, 25]$. Figure 4 shows the experimental results on ALOI and HW. We can observe that: (1) A relatively low number of experts tends to result in sub-optimal performance, because a limited number of experts may not fully exploit the diverse information inherent in multi-view data. (2) The marginal gain in performance diminishes as the number of experts increases beyond a certain threshold, which indicates that an optimal balance exists between model complexity and performance.

Impact of the Number of Selected Experts. Here we analyze the impact of the number of selected experts, *i.e.*, K, on clustering performance. Figure 5 illustrates the evaluation metrics' distributions on ALOI and HW, with K vary-



(a) ALOI (b) HW

Figure 5: Performance fluctuation under the different selected expert numbers.

ing within the range of $[1, 10]$. Note that the 'violin' may exceed the upper bound as we use kernel density estimation to approximate the performance distribution. We observe that a small K value, *e.g.*, 1, leads to large performance fluctuations, as the model may fail to fully leverage the diversity of the available experts. As the value of K increases, the model initially benefits from a broader integration of expert insights, enhancing the quality of the learned representations. Yet, further increasing K beyond a certain point results in diminishing returns or even negative influence. This may be due to the incorporation of redundant information from additional experts.

Ablation Study

To validate the necessity of each component in DMVC-CE, we conduct an ablation study to evaluate their influences. We construct several degradation models of DMVC-CE, i.e., **(1) ME (Multi-Extractor) Ensemble**: Replacing the MoE learner with multiple separate extractors. **(2) w/o** ℓ_{e}: The equilibrium loss is removed from the training process. **(3) w/o** $\mathcal{R}_{\mathcal{H}}^2$: The distinctiveness enhancer is excluded from the model. **(4) w/o** ℓ_{e} & $\mathcal{R}_{\mathcal{H}}^2$: Both distinctiveness enhancer and equilibrium loss are excluded. Table 3 shows the performance comparison of these degradation models with DMVC-CE, where we have the following observations. Firstly, the lack of direct collaboration in ME Ensemble results in underutilized cross-view correlations, and the design of view-specific feature extractors obviously limits their versatility. In contrast, our MoE architecture enables dynamic collaboration, with multi-view data processed by multiple experts shared across views, allowing for the capture of view-specific information while considering cross-view correlations. Secondly, the absence of the equilibrium loss ℓ_{e} leads to a certain performance decline, which demonstrates that the unbalanced expert selection limits the representation learning capability. Thirdly, the performance also degrades when the distinctiveness enhancer is removed. This reveals the crucial role of expert diversity in ensuring that each expert contributes unique and complementary information, and avoiding redundancy in the learned representations. Finally, the base model, where only the MoE representation learner remains, results in the most significant performance decline. Overall, the ablation study fully demonstrates the effectiveness of each component in DMVC-CE, which enables balanced dynamic collaboration among experts and fully exploits the diverse information within multi-view data.

	Metric	ALOI	HW
ME Ensemble	ACC	54.90±2.42	64.38±2.67
	NMI	56.06±3.71	65.61±1.31
	ARI	38.92±4.94	53.48±2.14
w/o ℓ_{e}	ACC	87.67±1.71	93.97±4.52
	NMI	85.29±1.79	94.05±3.88
	ARI	77.92±2.38	90.41±6.70
w/o $\mathcal{R}_{\mathcal{H}}^2$	ACC	84.02±4.19	93.37±5.10
	NMI	82.93±2.79	92.65±5.43
	ARI	73.64±4.68	89.03±8.12
w/o ℓ_{e} & $\mathcal{R}_{\mathcal{H}}^2$	ACC	80.82±5.43	87.00±9.52
	NMI	82.50±3.86	89.07±6.32
	ARI	70.14±7.38	82.24±10.61
DMVC-CE	ACC	**91.54±0.59**	**98.75±1.01**
	NMI	**88.41±0.50**	**98.00±1.32**
	ARI	**83.81±0.55**	**97.40±2.02**

Table 3: Ablation study results on ALOI and HW.

Conclusion

In this paper, we propose DMVC-CE, a simple yet effective deep multi-view clustering method that leverages the MoE paradigm to address the limitations of existing MVC frameworks. DMVC-CE enables collaboration across multiple experts to effectively capture the complementary information from multi-view data, where a gating network is introduced to dynamically select appropriate experts for representation learning. Additionally, we introduce an equilibrium loss and a distinctiveness enhancer to ensure the balance in expert selection and facilitate the diversity of experts. Empirical results on several benchmark datasets compared with state-of-the-art MVC baselines fully demonstrate the superiority of our method. Despite these advancements, one limitation of DMVC-CE lies in its two-stage framework, where the framework may benefit from incorporating a clustering objective to build an end-to-end model.

Acknowledgments

This research is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-018) and the Science and Technology Development Fund (FDCT), Macau SAR (file no. 0123/2023/RIA2, 001/2024/SKL).

References

Cai, H.; Hu, Y.; Qi, F.; Hu, B.; and Cheung, Y.-m. 2024a. Deep tensor spectral clustering network via ensemble of multiple affinity tensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cai, J.; Fan, J.; Guo, W.; Wang, S.; Zhang, Y.; and Zhang, Z. 2022a. Efficient deep embedded subspace clustering. In *CVPR*, 1–10.

Cai, J.; Wang, S.; Xu, C.; and Guo, W. 2022b. Unsupervised deep clustering via contractive feature representation and focal loss. *Pattern Recognition*, 123: 108386.

Cai, J.; Zhang, Y.; Fan, J.; Du, Y.; and Guo, W. 2024b. Dual contrastive graph-level clustering with multiple cluster perspectives alignment. In *IJCAI*, 3770–3779.

Cai, J.; Zhang, Y.; Wang, S.; Fan, J.; and Guo, W. 2024c. Wasserstein embedding learning for deep clustering: A generative approach. *IEEE Transactions on Multimedia*.

Chen, J.; Mao, H.; Woo, W. L.; and Peng, X. 2023a. Deep multiview clustering by contrasting cluster assignments. In *ICCV*, 16752–16761.

Chen, M.-S.; Huang, L.; Wang, C.-D.; and Huang, D. 2020. Multi-view clustering in latent embedding space. In *AAAI*, volume 34, 3513–3520.

Chen, M.-S.; Wang, C.-D.; Huang, D.; Lai, J.-H.; and Philip, S. Y. 2024. Concept factorization based multiview clustering for large-scale data. *IEEE Transactions on Knowledge and Data Engineering*.

Chen, M.-S.; Wang, C.-D.; Huang, D.; Lai, J.-H.; and Yu, P. S. 2022. Efficient orthogonal multi-view subspace clustering. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 127–135.

Chen, Z.; Fu, L.; Yao, J.; Guo, W.; Plant, C.; and Wang, S. 2023b. Learnable graph convolutional network and feature fusion for multi-view learning. *Information Fusion*, 95: 109–119.

Fang, U.; Li, M.; Li, J.; Gao, L.; Jia, T.; and Zhang, Y. 2023. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 12350–12368.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Fu, L.; Chen, Z.; Chen, Y.; and Wang, S. 2022. Unified low-rank tensor learning and spectral embedding for multi-view subspace clustering. *IEEE Transactions on Multimedia*, 25: 4972–4985.

He, M.; Lv, G.; He, W.; Fan, J.; and Zeng, G. 2021. DeepME: Deep mixture experts for large-scale image classification. In *IJCAI*, 722–728.

Houthuys, L.; Langone, R.; and Suykens, J. A. 2018. Multi-view kernel spectral clustering. *Information Fusion*, 44: 46–56.

Hu, Y.; Guo, E.; Xie, Z.; Liu, X.; and Cai, H. 2023. Robust multi-view clustering through partition integration on stiefel manifold. *IEEE Transactions on Knowledge and Data Engineering*.

Huang, S.; Kang, Z.; Tsang, I. W.; and Xu, Z. 2019. Auto-weighted multi-view clustering via kernelized graph learning. *Pattern Recognition*, 88: 174–184.

Huang, S.; Zhang, Y.; Fu, L.; and Wang, S. 2022. Learnable multi-view matrix factorization with graph embedding and flexible loss. *IEEE Transactions on Multimedia*, 25: 3259–3272.

Huang, W.; Yang, S.; and Cai, H. 2024. Generalized information-theoretic multi-view clustering. *Advances in Neural Information Processing Systems*, 36.

Li, J.; Gao, Q.; Wang, Q.; Yang, M.; and Xia, W. 2024. Orthogonal non-negative tensor factorization based multi-view clustering. *Advances in Neural Information Processing Systems*, 36.

Li, P.; Dau, H.; Puleo, G.; and Milenkovic, O. 2017. Motif clustering and overlapping clustering for social network analysis. In *Proceedings of the IEEE Conference on Computer Communications*, 1–9. IEEE.

Li, X.; Sun, Y.; Sun, Q.; and Ren, Z. 2023a. Enforced block diagonal graph learning for multikernel clustering. *IEEE Transactions on Computational Social Systems*, 11(2): 1753–1765.

Li, X.; Sun, Y.; Sun, Q.; Ren, Z.; and Sun, Y. 2023b. Cross-view graph matching guided anchor alignment for incomplete multi-view clustering. *Information Fusion*, 100: 101941.

Li, X.; and Xu, H. 2023. MEID: mixture-of-experts with internal distillation for long-tailed video recognition. In *AAAI*, volume 37, 1451–1459.

Li, Z.; Tang, C.; Liu, X.; Zheng, X.; Zhang, W.; and Zhu, E. 2021. Consensus graph learning for multi-view clustering. *IEEE Transactions on Multimedia*, 24: 2461–2472.

Li, Z.; Wang, Q.; Tao, Z.; Gao, Q.; Yang, Z.; et al. 2019. Deep adversarial multi-view clustering network. In *IJCAI*, volume 2, 4.

Liu, J.; Wang, C.; Gao, J.; and Han, J. 2013. Multi-view clustering via joint nonnegative matrix factorization. In *ICDM*, 252–260. SIAM.

Liu, S.; Liao, Q.; Wang, S.; Liu, X.; and Zhu, E. 2024. Robust and consistent anchor graph learning for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, Y.; Liang, K.; Xia, J.; Zhou, S.; Yang, X.; Liu, X.; and Li, S. Z. 2023a. Dink-net: Neural clustering on large graphs. In *ICML*, 21794–21812. PMLR.

Liu, Y.; Yang, X.; Zhou, S.; Liu, X.; Wang, Z.; Liang, K.; Tu, W.; Li, L.; Duan, J.; and Chen, C. 2023b. Hard sample aware network for contrastive deep graph clustering. In *AAAI*, volume 37, 8914–8922.

Luo, C.; Xu, J.; Ren, Y.; Ma, J.; and Zhu, X. 2024. Simple contrastive multi-view clustering with data-level fusion. In *IJCAI*, 4697–4705.

Nie, F.; Cai, G.; and Li, X. 2017. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, 2408–2414.

Pu, J.; Cui, C.; Chen, X.; Ren, Y.; Pu, X.; Hao, Z.; Philip, S. Y.; and He, L. 2024. Adaptive feature imputation with latent graph for deep incomplete multi-view clustering. In *AAAI*, volume 38, 14633–14641.

Qi, Z.; Meng, L.; He, W.; Zhang, R.; Wang, Y.; Qi, X.; and Meng, X. 2024. Cross-training with multi-view knowledge fusion for heterogenous federated learning. *arXiv preprint arXiv:2405.20046*.

Qi, Z.; Wang, Y.; Chen, Z.; Wang, R.; Meng, X.; and Meng, L. 2022. Clustering-based curriculum construction for sample-balanced federated learning. In *Proceedings of the CAAI International Conference on Artificial Intelligence*, 155–166. Springer.

Ren, Y.; Pu, J.; Yang, Z.; Xu, J.; Li, G.; Pu, X.; Philip, S. Y.; and He, L. 2024. Deep clustering: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Sadybekov, A. V.; and Katritch, V. 2023. Computational approaches streamlining drug discovery. *Nature*, 616(7958): 673–685.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of the International Conference on Learning Representations*.

Sun, M.; Zhang, P.; Wang, S.; Zhou, S.; Tu, W.; Liu, X.; Zhu, E.; and Wang, C. 2021. Scalable multi-view subspace clustering with unified anchors. In *Proceedings of the ACM International Conference on Multimedia*, 3528–3536.

Tan, Y.; Liu, Y.; Wu, H.; Lv, J.; and Huang, S. 2023. Metric multi-view graph clustering. In *AAAI*, volume 37, 9962–9970.

Tzortzis, G.; and Likas, A. 2012. Kernel-based weighted multi-view clustering. In *ICDM*, 675–684. IEEE.

Wan, X.; Liu, J.; Gan, X.; Liu, X.; Wang, S.; Wen, Y.; Wan, T.; and Zhu, E. 2024. One-step multi-view clustering with diverse representation. *IEEE Transactions on Neural Networks and Learning Systems*.

Wan, X.; Liu, X.; Liu, J.; Wang, S.; Wen, Y.; Liang, W.; Zhu, E.; Liu, Z.; and Zhou, L. 2023. Auto-weighted multi-view clustering for large-scale data. In *AAAI*, volume 37, 10078–10086.

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

Wen, Z.; Ling, Y.; Ren, Y.; Wu, T.; Chen, J.; Pu, X.; Hao, Z.; and He, L. 2024. Homophily-related: Adaptive hybrid graph filter for multi-View graph clustering. In *AAAI*, volume 38, 15841–15849.

Wu, Z.; Lin, X.; Lin, Z.; Chen, Z.; Bai, Y.; and Wang, S. 2023. Interpretable graph convolutional network for multi-view semi-supervised learning. *IEEE Transactions on Multimedia*, 25: 8593–8606.

Wu, Z.; Zhang, Z.; and Fan, J. 2024. Graph convolutional kernel machine versus graph convolutional networks. *Advances in Neural Information Processing Systems*, 36.

Xu, J.; Ren, Y.; Tang, H.; Yang, Z.; Pan, L.; Yang, Y.; Pu, X.; Philip, S. Y.; and He, L. 2022. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7): 7470–7482.

Xu, R.; and Wunsch, D. 2005. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3): 645–678.

Yang, B.; Zhang, X.; Nie, F.; Wang, F.; Yu, W.; and Wang, R. 2020. Fast multi-view clustering via nonnegative and orthogonal factorization. *IEEE Transactions on Image Processing*, 30: 2575–2586.

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

Yuksel, S. E.; Wilson, J. N.; and Gader, P. D. 2012. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8): 1177–1193.

Zhan, K.; Nie, F.; Wang, J.; and Yang, Y. 2019. Multiview consensus graph clustering. *IEEE Transactions on Image Processing*, 28(3): 1261–1270.

Zhang, C.; Jia, X.; Li, Z.; Chen, C.; and Li, H. 2024. Learning cluster-wise anchors for multi-view clustering. In *AAAI*, volume 38, 16696–16704.

Zhang, C.; Liu, Y.; and Fu, H. 2019. Ae2-nets: Autoencoder in autoencoder networks. In *CVPR*, 2577–2585.

Zhao, H.; Ding, Z.; and Fu, Y. 2017. Multi-view clustering via deep matrix factorization. In *AAAI*, volume 31.

Zhou, L.; Du, G.; Lü, K.; Wang, L.; and Du, J. 2024. A survey and an empirical evaluation of multi-view clustering approaches. *ACM Computing Surveys*, 56(7): 1–38.

Zhu, P.; Yao, X.; Wang, Y.; Hui, B.; Du, D.; and Hu, Q. 2024. Multiview deep subspace clustering networks. *IEEE Transactions on Cybernetics*.