

Article

MoE Based Consistency and Complementarity Mining for Multi-View Clustering

Xiaoping Wang ¹, Yang Cao ², Yifan Zhang ^{3,*}, Hanlu Ren ⁴ and Qiyue Yin ⁵¹ Yellow River Engineering Consulting Co., Ltd., Zhengzhou 450003, China; wang_xp@yrec.cn² Bei Fang Investigation, Design & Research Co., Ltd., Tianjin 300202, China; cao_y2@bidr.com.cn³ Future Education College, Qingdao Hengxing University of Science and Technology, Qingdao 266100, China⁴ Science and Technology Consultation Center of Haihe River Water Conservancy Commission, Ministry of Water Resources, Tianjin 300171, China; renhanlu@hwcc.gov.cn⁵ Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; qyyin@nlpr.ia.ac.cn

* Correspondence: zzyiff@126.com

Abstract

Multi-view clustering, which improves clustering performance by using the complementary and consistent information from multiple diverse feature sets, has been attracting increasing research attention owing to its broad applicability in real world scenarios. Conventional approaches typically leverage this complementarity by projecting different views into a common embedding space using view-specific or shared non-linear neural networks. This unified embedding is then fed into standard single-view clustering algorithms to obtain the final clustering results. However, a single common embedding may be insufficient to capture the distinct or even contradictory characteristics of multi-view data, due to the divergent representational capacities of different views. To address this issue, we propose a mixture of experts (MoE) based embedding learning method that adaptively models inter-view relationships. This architecture employs a typical MoE module as a projection layer across all views, which uses shared expert and several groups of experts for consistency and complementarity mining. Furthermore, a Kullback-Leibler divergence based objective with over clustering is designed for clustering-oriented embedding learning. Extensive experiments on six benchmark datasets confirm that our method achieves superior performance compared to a number of state-of-the-art approaches.

Keywords: multi-view clustering; MoE; over clustering; Kullback-Leibler divergence

1. Introduction

Current information is often characterized by multiple distinct feature representations. For example, hydrological data on rivers and lakes may integrate measurements from monitoring stations as well as remote sensing imagery. These diverse feature sets capture semantically related content while exhibiting complementary information and consistency across different views. Such a property forms the foundation of multi-view learning, which holds the potential to achieve better performance than single-view methods [1]. When multi-view learning is applied to clustering tasks, it gives rise to multi-view clustering an effective approach for organizing large-scale unlabeled multi-view data. This methodology has been widely adopted in various domains, including pattern recognition and data mining.

In recent years, a variety of multi-view clustering methods have been developed [2]. Among these, subspace-based approaches have attracted considerable attention by seeking a low-dimensional unified embedding space across different views [3]. To achieve



Received: 17 December 2025

Revised: 24 January 2026

Accepted: 3 February 2026

Published: 6 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

effective low-dimensional representation, canonical correlation analysis (CCA) methods learn view-specific projections that maximize cross-view correlations, thereby capturing inter-view relationships [4,5]. Spectral analysis methods utilize the spectral decomposition of Laplacian matrices constructed from all views to derive a unified embedding, preserving structural consistency across views [6,7]. Matrix factorization techniques decompose multiple high-dimensional feature matrices into base matrices and a common low-dimensional embedding, aiming to reveal latent structures indicative of cluster assignments [8,9]. Furthermore, all the aforementioned methods can be integrated with advanced deep neural networks to further enhance clustering performance [10–12].

Subspace-based multi-view clustering methods typically adopt two primary learning strategies to compute a unified embedding. The first strategy employs view-specific neural networks, which utilize distinct networks tailored to individual views to learn view-specific representations. To ensure consistency across views, a similarity-based loss is incorporated during training, encouraging alignment of the learned representations from different views. The second strategy utilizes shared neural networks, where a single network processes all views to learn a common representation. In this case, a reconstruction loss is commonly applied to preserve fidelity between the input and its reconstructed output, ensuring that the shared representation retains essential features of the original data. However, due to varying representation capacities and potential noise in individual views, existing methods that rely exclusively on either shared or view-specific networks often fail to fully capture the diverse characteristics of each view [13,14]. This limitation highlights the need for more advanced approaches that can effectively integrate the advantages of both strategies while addressing their respective shortcomings.

In response to the aforementioned limitations, inspired from the concept of mixture of experts (MoE) used in large models for efficient pretraining, we propose MoE based multi-view embedding learning (MEL-MoE), a novel framework designed to flexibly model multi-view relationships with enhanced versatility. Specifically, the method features a typical MoE module as a projection layer for all the views, as shown in Figure 1. To ensure the equilibrium in the expert selection, and facilitate the complementarity mining among multiple views, several groups of experts are designed and selected based on routers driven by each view. Furthermore, a shared expert is designed in each layer as a base to prompt consistency between different views. The learning process is regularized by three objectives: a reconstruction loss to ensure input fidelity, a similarity loss to align the view-specific representations, and a Kullback-Leibler (KL) divergence term with over clustering that serves as a clustering constraint. Extensive experiments on benchmark datasets demonstrate the superior performance of MEL-MoE over state-of-the-art baselines.

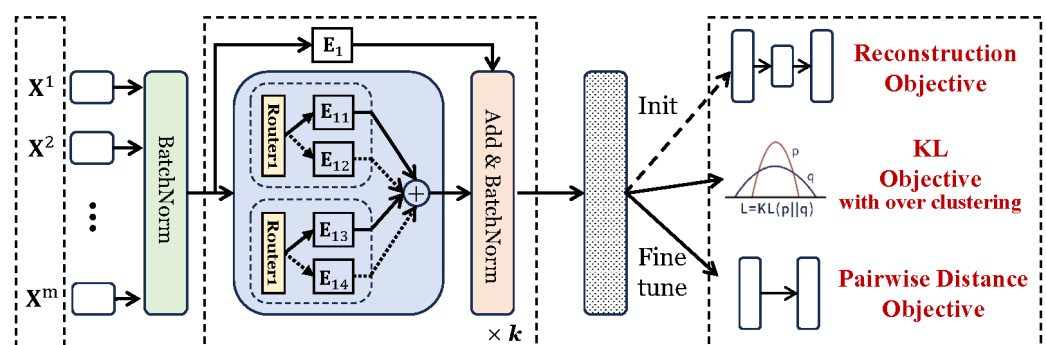


Figure 1. Framework of the proposed MEL-MoE. Each layer of MEL-MoE comprises two synergistic modules: a shared expert network and routing network augmented expert groups. A tailored objective function paired with the above layer design enables the efficient extraction of complementary and consistent information from multi-view data.

The main contributions of this work are summarized as follows:

- We propose a MoE based multi-view embedding learning method that effectively leverages inter-view consistency and complementarity, enabling adaptive modeling of the complex relationships in multi-view data.
- Through a rigorous evaluation on multiple benchmark datasets, our approach shows substantial performance improvements over existing state-of-the-art methods.

This paper is structured as follows: Section 2 is devoted to a review of related work on multi-view clustering. Our proposed methodology is then introduced in Section 3. Section 4 provides a comprehensive presentation of the experimental results and discussion. Finally, conclusions and future work are outlined in Section 5.

2. Related Work

2.1. Multi-View Clustering

To date, a number of multi-view clustering approaches have been developed, which can be broadly categorized into four types, with promising results demonstrated across various applications [7,15–17].

Graph-based methods focus on learning a unified similarity matrix, which then serves as the input for spectral clustering to obtain the final results [18–23]. A prominent line of these methods employ subspace segmentation based on sparse or low-rank self-representation, which have gained considerable popularity in multi-view clustering due to its encouraging outcomes. Co-training-based approaches [24], on the other hand, operate within a framework where the clustering of one view is consistently refined using information from other views, leading to enhanced overall performance. For example, Kumar et al. [25] applied this idea to guide the construction of the Laplacian matrix across views, an approach later extended by Liu [26] to handle large-scale datasets. In contrast, late fusion methods concentrate on aggregating or aligning clustering results obtained from individual views to achieve a robust consensus [27,28]. Commonly used techniques in this category include probabilistic models [29] and matrix optimization strategies [30]. It is also worth noting that replacing traditional linear mappings with non-linear neural networks has recently led to significant performance improvements in multi-view clustering [31–33].

The last category of methods are subspace based, which learn a unified embedding for single view clustering. Kumar et al. [6] enforce consistency across views by constraining the spectral embeddings to be similar to each other or to a common unified representation. Andrew et al. [5] utilize deep neural networks to learn embeddings that maximize cross-view correlation. Zhao et al. [12] propose a deep matrix factorization framework to obtain a unified embedding. Yin et al. [13] assume the presence of a latent unified embedding, from which view-specific embeddings are generated to infer data similarities. Li et al. [34] develop a deep adversarial clustering network to learn multiple embeddings in a coordinated manner. Overall, subspace-based multi-view clustering methods typically adopt one of two neural network architectures to leverage both the complementarity and consistency among views. The first strategy utilizes view-specific networks to extract distinctive representations from each view, while employing a similarity-based loss function to maintain cross-view consistency. The second strategy employs a shared network to learn a consensus representation common to all views, usually through a reconstruction loss that preserves essential information from the original inputs.

However, we contend that existing methods often fail to sufficiently capture the intricate interdependencies among multiple views, a capability crucial for handling the diverse yet interrelated nature of multi-view data. To bridge this gap, this paper introduces an enhanced approach to alleviate this problem. It should be noted that Wang et al. [14] proposed a view-driven modular network, which is designed to capture the unique or

conflicting attributes of multi-view data. In contrast to this approach, our method adopts a substantially different architectural design, such as using shared and groups of experts to replace the soft connected fully connected neural network (FCNN), enabling it to better model the characteristics of multi-view data. Furthermore, our method achieves superior performance over the aforementioned approach, with detailed comparative results presented in the Experiments section.

2.2. Mixture of Experts

The concept of MoE can be traced back to ensemble learning, as both integrate multiple models, yet they differ significantly in implementation [35]. The MoE architecture consists of multiple specialized sub-models (the experts), each contributing within its domain of expertise. A mechanism known as the gating network determines which expert(s) are activated for a given input. As a powerful model architecture, MoE saw significant advancements in 2017, when a team led by Quoc Le, Geoffrey Hinton, and Jeff Dean introduced a novel MoE layer that dramatically enhanced the scale and efficiency of large models by incorporating sparsity [36]. Currently, MoE is emerging as a prominent topic in the era of large-scale models [37].

Recently, a few researchers have tried to introduce MoE in multi-view clustering. Zhang et al. [38] proposed MoE as representation learner for multi-view clustering, which uses a equilibrium loss and distinctiveness loss for expert selection. Compared with the above approach, our method introduces groups of experts and shared experts in each layer, which is more flexible and effective. Overall, the potential of MoE to address challenges in multi-view clustering has yet to be fully investigated.

3. MEL-MoE Method

In this paper, matrices and vectors are denoted by bold uppercase and lowercase letters, respectively. Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ represent a dataset of n data samples, each described by V distinct views. The v -th view is denoted as $\mathbf{X}^v = \{\mathbf{x}_i^v\}_{i=1}^n \in \mathbb{R}^{n \times d_v}$, where d_v is the feature dimensionality of that view. Our goal is to partition the n samples into c clusters. Note that determining c is beyond the scope of this work. Unless stated otherwise, superscripts and subscripts are used to index views and individual data elements, respectively.

The overall framework is shown in Figure 1. Unlike conventional multi-view clustering methods, which employ FCNN in each layer, we adopt a specially designed MoE network for every layer of our model. Specifically, each layer consists of an shared expert network, several groups of specific expert network, and corresponding router network. The shared expert network is designed as a foundational component to ensure cross-view consistency, while groups of view-specific expert networks are selected for different views to capture distinctive view-specific characteristics.

As shown in Figure 1, Data from each view is processed by the shared expert network and the selected groups of view-specific expert networks, followed by a summation operation. Mathematically, each layer can be formalized as:

$$\mathbf{Z}_k^v = \underbrace{f(\mathbf{Z}_{k-1}^v; \phi_k)}_{\text{Shared expert}} + \underbrace{\sum_{m=1}^M \sum_{t=1}^2 \pi_t(\mathbf{Z}_{k-1}^v; \phi_{kmt}) \odot f(\mathbf{Z}_{k-1}^v; \phi_{kmt})}_{\text{Groups of view - specific expert}} \quad (1)$$

where $\mathbf{Z}_k^v$ represents the calculated embedding for view v in the k -th layer. $f(\cdot)$ is the mapping function consisting of a fully connected neural network layer, with ϕ_k being the shared expert parameters in the k -th layer for all views, ϕ_{kmt} being the t -th expert parameters of the m -th group in the k -th layer for all the views. π is the routing function, with θ_{km} being the routing parameters of the m -th group in the k -th layer for all the views,

and π_t is the t -th value of π . Generally, π is a fully connected neural network layer with Softmax operation. By using the routing function, the most relevant experts are selected.

Since groups of experts are used to maintain equilibrium for the expert selection, no other constraints are used. Following previous study, we use top-1 expert selection in the framework. Accordingly, π_t is calculated as 0 or 1 based on the routing function.

To this end, we design the following objective to learn the above embedding, hoping the diverse and complementary features can be well mined:

$$\mathcal{L} = \sum_{v=1}^V \mathcal{L}_{recon}^v + \alpha \mathcal{L}_{pair} + \beta \mathcal{L}_{klo} \quad (2)$$

where $\mathcal{L}_{recon}^v$ is reconstruction loss of the v -th view for learning representation $\mathbf{Z}_k^v$, $\mathcal{L}_{pair}$ is a consistency loss function to constrain different $\mathbf{Z}_k^v$ to be similar, hoping the consistency is preserved, $\mathcal{L}_{klo}$ is a Kullback-Leibler divergence loss function to boost the clustering embedding obtained from all $\mathbf{Z}_k^v$, hoping clustering-oriented embedding is learned with the concept of over clustering.

3.1. Reconstruction Loss $\mathcal{L}_{recon}^v$

With Equation (1), we can learn the k -th layer representation of view v . Suppose there are l layers of the encoder (Parametrize as ϕ), the learned embedding for view v is obtained as $\mathbf{Z}_l^v$. We use the same network setting (Parametrize as θ) as the decoder of each view, then the reconstruction of view v can be defined as $\tilde{\mathbf{X}}^v$. Finally, the reconstruction loss can be represented as:

$$\mathcal{L}_{recon}^v = \|\mathbf{X}^v - \tilde{\mathbf{X}}^v\|_F^2 \quad (3)$$

Note that when constructing the decoder network for view v , the routing weights are not independently learned; instead, they are directly mirrored from the weights of the routing network in the encoder to reduce training complexity. Besides, to balance the expert selection, in a batch, data are evenly sampled from different views.

3.2. Consistency Loss $\mathcal{L}_{pair}$

After learning the embeddings $\{\mathbf{Z}_l^v, v = 1, \dots, V\}$ of different views, a pairwise-based objective is used to force different embeddings to be similar. The loss is defined as:

$$\mathcal{L}_{pair} = \sum_{v=1}^V \sum_{u=v+1}^V \|\mathbf{Z}_l^v - \mathbf{Z}_l^u\|_F^2 \quad (4)$$

With this loss, consistency between different views can be guaranteed, which is a key point for mining multi-view characteristics.

3.3. Kullback-Leibler Divergence Loss $\mathcal{L}_{klo}$

With the learned embeddings for all views, a unified embedding can be obtained by using an aggregation function, as formalized by:

$$\mathbf{Z} = [\mathbf{Z}_l^1, \mathbf{Z}_l^2, \dots, \mathbf{Z}_l^V] \quad (5)$$

By using the concatenation strategy, information for different views are preserved, which enhances the complementarity mining of multi-view data.

With $\mathbf{Z}_l$, Clustering is typically performed by applying typical single-view algorithms. However, this approach decouples the clustering objective from the embedding learning process, which can yield suboptimal embeddings that lack the specific characteristics required for effective clustering and thus lead to diminished performance. To solve this problem, we use a Kullback-Leibler divergence based objective to boost the learned unified embedding. Specifically, the process begins with the unified embeddings, from which

initial cluster centroids and soft assignments (i.e., the posterior probability of a sample belonging to each cluster) are first estimated. These initial estimates enable the construction of auxiliary objectives to refine the assignments based on principles like boosting high-confidence predictions.

Computationally, this involves two steps: First, the cluster centroids $\mathbf{v}_j, j = 1, \dots, \gamma c$ are calculated from the initial embeddings, usually with K-means. Then, the cluster assignment for each sample is computed by applying a normalization operator (e.g., softmax) to the standardized distances between the sample and all centroids. The assignment probability is formally defined as:

$$q_{ij} = \frac{(1 + \|\mathbf{z}_i - \mathbf{v}_j\|^2)^{-1}}{\sum_j (1 + \|\mathbf{z}_i - \mathbf{v}_j\|^2)^{-1}}, \forall i = 1, \dots, n, j = 1, \dots, \gamma c \quad (6)$$

where q_{ij} is the probability sample i belongs to cluster j .

Note that the number of cluster centroids is set to be γ times of the real number. The over-clustering strategy is employed to enforce a more fine-grained data partition, thereby enhancing both cluster purity and feature quality. This approach aids the neural network in determining how to partition the data effectively, ultimately improving clustering performance, which is widely used in data mining field [39,40].

Then we can define $\mathcal{L}_{klo}$ as:

$$\mathcal{L}_{klo} = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (7)$$

where p_{ij} is an auxiliary variable calculated as:

$$p_{ij} = (q_{ij})^2 \quad (8)$$

which is then normalized as:

$$p_{ij} = \frac{p_{ij}}{\sum_j p_{ij}} \quad (9)$$

3.4. Optimization

We optimize Equation (1) with two steps. Firstly, we optimize $\mathcal{L}_{recon}^v$ (Equation (3)) to obtain initial embeddings of $\{\mathbf{Z}_l^v, v = 1, \dots, V\}$, where the encoder network and the decoder network parameters are updated.

Secondly, we optimize $\mathcal{L}$ (Equation (2)) without $\mathcal{L}_{recon}^v$ using backpropagation, where the encoder network and the clustering centroids are updated. The gradients calculated are:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}_j} = - \sum_i \left(1 + \|\mathbf{z}_i - \mathbf{v}_j\|^2\right)^{-1} \times (p_{ij} - q_{ij}) (\mathbf{z}_i - \mathbf{v}_j) \quad (10)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}_i} = - \sum_j \left(1 + \|\mathbf{z}_i - \mathbf{v}_j\|^2\right)^{-1} \times (p_{ij} - q_{ij}) (\mathbf{z}_i - \mathbf{v}_j) \quad (11)$$

Finally, we obtain the clustering results by using K-means on $\mathbf{Z}$, as shown in Algorithm 1.

Convergence analysis: In the first step, we employ a standard backpropagation algorithm, which ensures the convergence of the model. In the second step, a similar procedure is adopted, leveraging the gradient backpropagation derived from Equations (10) and (11), a strategy that also guarantees convergence at this stage.

Algorithm 1: Our proposed MEL-MoE**Input:** $\mathbf{X}$ with c predefined clusters.**Output:** Cluster assignments for each $\mathbf{x}_i$.

1. Initialize the parameters ϕ and θ , respectively;
2. Optimize the parameters ϕ and θ based on the objective function $\mathcal{L}_{recon}^v$ defined in Equation (3);
3. Calculate the clustering centroids $\mathbf{v}_j, j = 1, \dots, \gamma c$, using the K-means algorithm;
4. Optimize the parameters ϕ and $\mathbf{v}_j$ based on the objective function defined in Equation (2) excluding $\mathcal{L}_{recon}^v$;
5. Return clustering label for all the samples using K-means on $\mathbf{Z}$.

4. Experiments

4.1. Experimental Datasets

Six widely used multi-view clustering datasets are used for performance verification.

Notting Hill (NH) is a video face dataset constructed from the movie Notting Hill. It contains 4660 images spanning 5 categories. Following the protocol in [12], we extract three feature types: intensity, LBP (Local Binary Patterns), and Gabor.

USPS (<http://archive.ics.uci.edu/ml/datasets/Multiple+Features>, accessed on 2 February 2026) is a dataset of handwritten digits sourced from Dutch utility maps. It includes 2000 samples distributed across 10 categories. In accordance with [6], we use two feature representations: 76 Fourier coefficients of the character contours and 216 profile correlations.

3Sources (<http://mlg.ucd.ie/datasets/3sources.html>, accessed on 2 February 2026) contains news articles collected from three online outlets: BBC, Reuters, and The Guardian. The subset used here consists of 169 stories covering 5 topics, with each story reported by all three sources. Each source is treated as a separate view, represented by word frequency features.

Cora (<http://lig-membres.imag.fr/grimal/data.html>, accessed on 2 February 2026) is an academic paper dataset covering domains such as neural networks, theory, and genetic algorithms. It contains 2708 papers categorized into 7 classes. We adopt two views: the citation links between documents and the word vectors derived from their content.

BBC (<http://mlg.ucd.ie/datasets/segment.html>, accessed on 2 February 2026) is a text dataset derived from the BBC and BBC Sport news corpora. It comprises 2012 documents divided into 5 thematic categories. We utilize two segment-based feature representations with dimensionalities of 6838 and 6790, respectively.

CCV (<http://www.ee.columbia.edu/ln/dvmm/CCV/>, accessed on 2 February 2026) is a video dataset containing 9317 YouTube videos labeled with 20 semantic categories. We employ two visual feature types: Scale-Invariant Feature Transform (SIFT) and Space-Time Interest Points (STIP).

4.2. Experimental Settings

Our experimental evaluation includes a comparison of our approach with several established multi-view clustering methods to thoroughly evaluate its performance.

SingleB: Spectral clustering is applied to each view individually, and the view yielding the best result is selected.

CCA [4]: A low-dimensional embedding is derived via Canonical Correlation Analysis on the multiple views, followed by K-means clustering.

Co-Pair, Co-Cent [6]: These methods regularize the spectral embeddings across views using pairwise constraints to enforce similarity between them, and centroid constraints to align them with a common latent embedding.

MultiDMF [12]: This method employs matrix factorization with deep neural networks to learn hierarchical base matrices and a unified representation for clustering.

MultiTE [13]: It jointly learns a unified embedding and view-specific embeddings, connected via a mapping matrix. K-means is then applied to the unified embedding for clustering.

MVCF [41]: The cluster indicator is directly optimized via a regression-based matrix factorization framework.

SGF, DGF [19]: These methods optimize similarity and dissimilarity graphs by formulating a unified objective that captures multi-view consistency and inconsistency.

SUMVC [42]: It employs the Bayes error rate to analyze clustering performance, guided by a proposed sufficient representation lower bound.

APADC [43]: An imputation-free deep learning method for incomplete multi-view clustering that aligns distributions during representation learning.

VDMN [14]: A view-driven modular network, which adaptively model the relationship between different views for complementarity and consistency mining.

MEL-MoE: The proposed MoE based multi-view embedding learning for multi-view clustering. Our model uses an encoder-decoder architecture analogous to APADC and VDMN, with differences limited to the objective function and layer definitions. Given the same parameter count, their computational complexity is comparable. For example, with an identical number of layers, the total parameters of multi-expert FCNNs per layer in our method can match the single FCNN per layer in APADC, and the multi-modular per layer in VDMN.

For Co-Reg-Pairwise, Co-Reg-Centroid, MultiDMF, MVCF, SGF, and DGF, we used the authors' publicly available implementations, carefully tuning the hyperparameters to optimize performance. The CCA baseline was implemented using the LSCCA package, where we adjusted the embedding dimensionality and regularization weight for optimal results. For MultiTE, SUMVC, APADC and VDMN, we developed our own implementations in accordance with the specifications in their respective publications.

For our method, a two-layer architecture was adopted across all experiments, where each layer comprises two groups of experts, with each group containing two experts. Every expert is implemented as a two-layer FCNN with 64 output nodes. The routing network is also configured as a two-layer FCNN, with 2 output nodes followed by softmax normalization. Since no reconstruction loss is employed in the second optimization stage, the hyperparameter α was set to 1. For the over-clustering hyperparameter γ , a fixed value of 2 was adopted. Meanwhile, the hyperparameter β was tuned over the set {0.001, 0.01, 0.1, 1, 10} to achieve optimal performance on each individual dataset. Before optimization, all the input features are normalized with L2-norm. For optimization, we employed the Adam optimizer, with the mini-batch size and learning rate set to 32 and 10^{-3} , respectively. In the first optimization stage, the training was conducted for 3000 epochs. Before the second optimization stage, K-means from Sklearn is used with default configuration, i.e., `init = 'k-means++'`, `n_init = 'auto'`, `max_iter = 300`, `tol = 0.0001`, `verbose = 0`, `random_state = None`, `copy_x = True`, `algorithm = 'lloyd'`. When performing the second optimization stage, the training process terminated when the proportion of samples with changed cluster assignments between two consecutive iterations fell below 1%. To obtain the final clustering labels, K-means is used with the same setting as the previous stage. Finally, steps 3–5 in Algorithm 1 were conducted with 20 independent runs, and we report the corresponding mean and standard deviation for the experimental results.

The clustering performance was assessed using two standard metrics: Accuracy (ACC) and Normalized Mutual Information (NMI). Their definitions are provided as follows:

$$\text{ACC} = \frac{\sum_{i=1}^n \delta(y_i, \text{map}(\tilde{y}_i))}{n} \quad (12)$$

$$\text{NMI} = \frac{I(y; \tilde{y})}{\sqrt{H(y)H(\tilde{y})}} \quad (13)$$

where y and $\tilde{y}$ indicate the true and the predicted labels, respectively. The $\text{map}()$ function applies the Hungarian algorithm to find the optimal permutation for matching predicted labels to the true labels. The indicator function $\delta(\cdot)$ returns 1 for identical inputs and 0 otherwise. Mutual information $I(\cdot)$ is computed using the entropy $H(\cdot)$, with higher ACC and NMI values indicating superior clustering results.

4.3. Performance of MEL-MoE

The clustering performance, as evaluated by ACC and NMI on the six benchmark datasets, is comprehensively summarized in Tables 1 and 2, respectively. The results indicate that our method achieves superior ACC and NMI over baseline algorithms on the vast majority of datasets.

Table 1. Experimental results in terms of ACC on the six datasets are presented below, with both the mean and standard deviation reported.

Dataset	NH	CCV	Cora	BBC	USPS	Source
SingleB	81.49 (7.80)	21.27 (0.67)	32.75 (1.47)	81.96 (5.57)	61.81 (4.81)	52.93 (3.59)
CCA	80.41 (8.06)	26.86 (2.23)	26.95 (1.58)	35.55 (8.49)	74.53 (4.96)	62.37 (6.76)
Co-Pair	78.17 (0.35)	23.49 (0.50)	44.57 (3.57)	85.85 (9.00)	75.85 (5.89)	58.37 (3.28)
Co-Cent	81.05 (0.56)	25.39 (0.93)	43.26 (2.98)	87.29 (7.99)	77.73 (5.89)	58.93 (3.07)
MultiDMF	87.10 (0.90)	23.87 (0.07)	50.19 (0.02)	88.25 (0.05)	83.30 (0.15)	71.42 (0.48)
MultiTE	82.25 (5.70)	26.88 (0.61)	59.95 (0.86)	93.74 (0.99)	85.96 (1.45)	82.91 (2.64)
MVCF	88.90 (0.35)	24.40 (0.35)	38.60 (0.23)	90.25 (0.15)	63.00 (1.37)	75.70 (4.31)
SGF	85.21 (0.30)	21.99 (0.61)	55.17 (0.00)	94.14 (0.00)	94.25 (0.00)	83.99 (0.13)
DGF	87.38 (0.23)	22.00 (0.00)	55.17 (0.03)	94.53 (0.00)	67.55 (5.64)	88.11 (0.26)
SUMVC	90.35 (0.05)	27.10 (0.35)	58.17 (0.05)	95.13 (0.08)	93.55 (0.64)	89.11 (0.16)
APADC	89.32 (0.15)	26.95 (0.12)	57.85 (0.13)	95.12 (0.18)	93.25 (0.56)	88.85 (0.14)
VDMN	92.78 (0.36)	28.17 (0.69)	61.85 (0.80)	96.34 (0.26)	95.10 (0.12)	90.12 (2.56)
MEL-MoE	93.15(0.24)	29.68 (0.43)	63.46 (0.65)	95.85 (0.12)	96.18 (0.24)	91.48 (1.24)

While Co-Pair and Co-Cent integrate multi-view data by applying regularization to spectral embeddings, methods like MultiDMF and MVCF primarily rely on matrix factorization. The latter employ view-specific mappings and similarity-based losses in the embedding space. In comparison, our approach introduces MoE mechanism across all views. This architecture is explicitly designed to model both the consistency and complementarity among distinct feature sets. Furthermore, the use of non-linear mappings enhances representation learning. These factors collectively contribute to the superior performance of our model.

While MultiTE posits that view-specific embeddings are generated from a latent unified embedding, our method learns them directly from the input features via a combina-

tion of MoE modular. This approach can yield more accurate and detailed view-specific representations. Similarly, while SGF and DGF explicitly model view consistency and inconsistency using similarity and dissimilarity graphs, our framework integrates the learning of complementarity and consistency directly into the representation learning process. This integrated approach enables a more effective derivation of the final representation, thereby improving clustering performance.

Table 2. Experimental results in terms of NMI on the six datasets are presented below, with both the mean and standard deviation reported.

Dataset	NH	CCV	Cora	BBC	USPS	Source
SingleB	70.48 (2.27)	19.03 (0.40)	18.20 (1.00)	62.32 (3.16)	59.12 (1.95)	53.38 (2.12)
CCA	77.05 (6.62)	22.71 (1.38)	1.30 (0.46)	17.14 (8.54)	75.54 (3.07)	60.39 (6.94)
Co-Pair	66.27 (0.27)	19.71 (0.38)	27.94 (2.00)	73.37 (4.30)	71.16 (1.45)	62.25 (2.76)
Co-Cent	76.71 (0.48)	22.09 (0.62)	24.47 (1.48)	73.39 (3.78)	73.40 (2.42)	62.25 (2.51)
MultiDMF	79.70 (0.50)	22.77 (0.08)	30.67 (0.02)	78.44 (0.07)	77.19 (0.08)	54.15 (0.21)
MultiTE	74.00 (2.87)	22.24 (0.26)	39.75 (0.25)	81.58 (0.60)	82.32 (0.53)	79.36 (1.95)
MVCF	76.40 (0.21)	22.50 (0.46)	18.10 (0.24)	80.12 (0.05)	63.30 (0.78)	65.10 (3.38)
SGF	86.28 (0.10)	23.34 (0.28)	42.48 (0.13)	82.48 (0.00)	88.94 (0.00)	77.23 (0.00)
DGF	86.24 (0.20)	24.00 (0.10)	45.34 (0.00)	82.71 (0.00)	78.53 (3.00)	76.32 (0.18)
SUMVC	88.28 (0.24)	25.10 (0.20)	44.86 (0.05)	82.15 (0.15)	86.54 (1.15)	78.52 (0.08)
APADC	88.15 (0.15)	25.06 (0.14)	44.95 (0.16)	82.05 (0.25)	86.25 (8.86)	78.84 (0.13)
VDMN	90.23 (1.47)	26.15 (0.46)	46.08 (0.38)	84.85 (0.63)	86.68 (0.23)	81.23 (1.64)
MEL-MoE	91.69 (0.34)	26.24 (0.37)	47.95 (0.46)	86.16 (0.42)	87.13 (0.18)	82.36 (0.96)

While SUMVC introduces a sufficient representation lower bound to minimize redundancy and learn consistent information across views, and APADC aligns feature distributions by minimizing inter-view mean discrepancy, our method explicitly models complementarity during representation learning. This is achieved through a dedicated architecture of shared and groups of expert layers, which effectively captures both shared and unique information from all views, thereby enhancing clustering performance.

Compared with VDMN, where a base network is used for all views and view-specific routing network is constructed for different views, our method employs MoE as embedding learning architecture, and over clustering is used for further enhancing the clustering representation. These two factors largely promote the representation ability of the unified embedding, thereby enhancing clustering performance.

To assess the clustering suitability of the learned unified embedding, we visualize it using t-SNE (with configuration being: perplexity = 30, iterations = 1000, initialisation = ‘pca’), as shown in Figure 2. The result demonstrates that our method produces more compact and well-separated clusters. This well-defined structure offers strong qualitative evidence that our embedding effectively captures the underlying data distribution, thereby leading to improved clustering performance.

4.4. Ablation Experiment

The MoE layer is a cornerstone of our proposed method. To evaluate its contribution, we conducted ablation studies on the layer number and the groups number selection. The baseline model MEL-MoE, employs a two-layer network, where each layer contains

two groups of experts with each group contain two experts. We systematically adjusted the network's depth and width to isolate and quantify their impact on overall performance.

- MEL-MoE-Layer-1: We reduce the MoE network depth to a single layer.
- MEL-MoE-Layer-4: We extend the MoE network depth to four layers.
- MEL-MoE-Groups-1: We reduce the groups number in each MoE layer to a single group.
- MEL-MoE-Groups-4: We extend the groups number in each MoE layer to four groups.
- MEL-MoE-Expert-1: We reduce the number of experts in each group to a single expert, which degenerate into a traditional fully connected neural network.
- MEL-MoE-Expert-4: We extend the number of experts in each group to four experts.

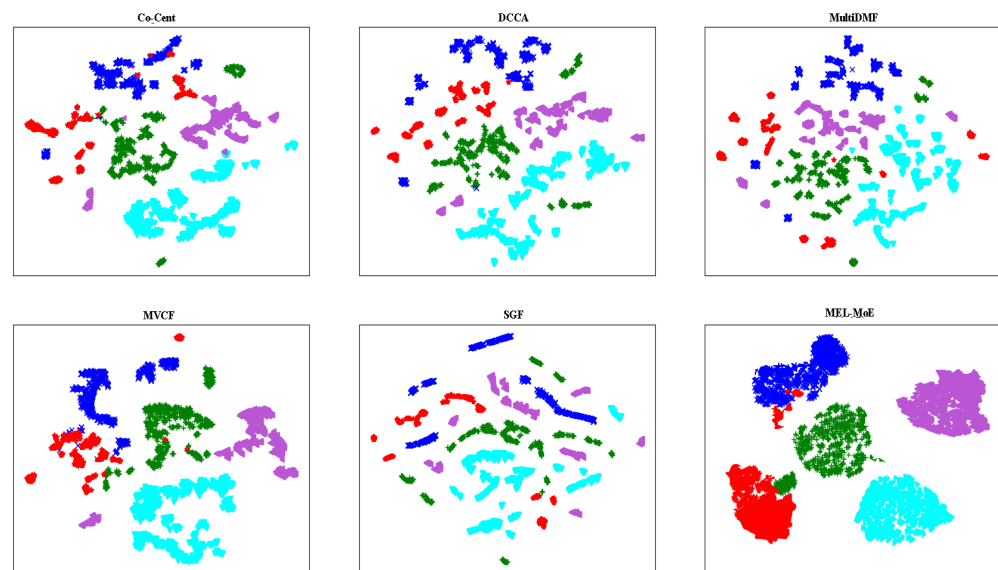


Figure 2. T-SNE for different methods on the NH dataset.

Table 3 summarizes the results on the NH dataset, which are separated into three groups of comparison, highlighting the impact of expert number, groups number and the layer number in the proposed method.

Table 3. Ablation study results on the NH dataset.

Methods	Accuracy	NMI
MEL-MoE-Layer-1	90.85 (0.23)	88.37 (0.32)
MEL-MoE-Layer-4	93.05 (0.19)	90.89 (0.42)
MEL-MoE	93.15 (0.24)	91.69 (0.34)
MEL-MoE-Groups-1	92.86 (0.21)	90.33 (0.35)
MEL-MoE-Groups-4	93.12 (0.25)	91.48 (0.16)
MEL-MoE	93.15 (0.24)	91.69 (0.34)
MEL-MoE-Expert-1	87.35 (0.68)	86.57 (0.92)
MEL-MoE-Expert-4	92.95 (0.34)	91.87 (0.12)
MEL-MoE	93.15 (0.24)	91.69 (0.34)

Fewer layers (MEL-MoE-Layer-1) than the baseline configuration limit the network's ability to learn rich representations, resulting in significantly degraded clustering performance. Although expanding the network to four layers (MEL-MoE-Layer-4) improves its feature learning capacity (thereby potentially enhancing clustering outcomes), the greater depth also increases training complexity. This heightened complexity can introduce drawbacks such as slower convergence or convergence to sub-optimal solutions.

Similar with the results of MEL-MoE-Layer-1 and MEL-MoE-Layer-4, a balanced number of groups are suitable for improving the clustering performance, which is also fit for the number of experts in each group. Future work will include similar ablation studies on additional datasets to validate these findings and to further assess the method's robustness across diverse data environments.

4.5. Hyperparameter Experiment

In the proposed method, apart from the hyperparameters used to balance different loss functions, there is a hyperparameter γ used to decide the degree of over clustering (as shown in Equation (6)). In this section, we investigate how the hyperparameter influence the final performance by gradually changing its value. The results in terms of Acc and NMI on the NH dataset is shown in Table 4.

From the results, an excessively small hyperparameter will result in indistinguishable categories, making it difficult to achieve optimal clustering results. Conversely, an overly large hyperparameter forces the separation of inherently indistinguishable categories, which likewise undermines the clustering structure. In summary, for hyperparameter selection, favorable experimental results can be achieved within the reasonable range of 1–2, thus circumventing suboptimal outcomes caused by excessively small or large hyperparameter values.

Beyond the quantitative experimental results, we provide a brief explanation to demonstrate the advantages of using over clustering. Consider a dataset consisting of two primary classes, each containing two subcategories. When directly partitioning the data into two classes, the inter-class distance between certain data points may be smaller than their intra-class distance, rendering them indistinguishable via direct class-level clustering. In contrast, learning to first partition the data into four subcategories necessitates the acquisition of more precise data representations to minimize the objective loss. In this scenario, subsequent aggregation of these four subcategories into the original two classes may be accomplished with greater ease and accuracy. Moving forward, we plan to conduct dedicated experiments in the future to empirically validate this hypothesized advantage.

Table 4. Hyperparameter study results on the NH dataset.

γ	0.5	1	1.5	2	5
Accuracy	83.16 (0.36)	92.15 (0.12)	92.96 (0.16)	93.15 (0.24)	91.68 (0.15)
NMI	81.68 (0.24)	90.85 (0.31)	91.35 (0.25)	91.69 (0.34)	90.12 (0.20)

5. Conclusions

We present a novel multi-view clustering method grounded in a MoE architecture. The proposed framework integrates a shared expert and several groups of experts as a module of MoE layer, to jointly exploit inter-view complementarity and consistency. Furthermore, a clustering-oriented constraint, implemented via a over clustering based KL divergence-based objective function, is incorporated to guide the learning of clustering-friendly representations. The empirical evaluation on six widely-used benchmarks provides compelling evidence for the superiority of our method over existing state-of-the-art approaches, underscoring its capability to distill shared and unique informational patterns across views.

Currently, our method is built upon the a priori known number of clusters, and it has several additional limitations that include potential routing instability and scalability trade-offs. Key future directions include automatic calculation of the number of clusters, validating robustness through extended experiments and adapting the method for incomplete and noisy multi-view clustering to solve practical challenge.

Author Contributions: Conceptualization, X.W. and Q.Y.; methodology, Y.Z. and Q.Y.; software, X.W. and Y.C.; validation, X.W., Y.C. and H.R.; formal analysis, X.W.; investigation, Y.Z. and H.R.; resources, Y.Z. and Q.Y.; data curation, X.W. and Y.C.; writing—original draft preparation, X.W.; writing—review and editing, Y.Z. and Q.Y.; visualization, Y.C.; supervision, Y.Z. and Q.Y.; funding acquisition, Q.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Key Research Project of Chinese Academy of Sciences (No. RCJJ-145-24-15).

Data Availability Statement: All datasets used are publicly available for download, see the Experimental datasets subsection for details.

Conflicts of Interest: Author Xiaoping Wang was employed by the company Yellow River Engineering Consulting Co., Ltd.; Author Yang Cao was employed by the company Bei Fang Investigation, Design & Research Co., Ltd.; Author Hanlu Ren was employed by the company Science and Technology Consultation Center of Haihe River Water Conservancy Commission, Ministry of Water Resources. The authors declare no conflicts of interest.

References

1. Fang, U.; Li, M.; Li, J.; Gao, L.; Jia, T.; Zhang, Y. A comprehensive survey on multi-view clustering. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 12350–12368. [\[CrossRef\]](#)
2. Zhou, L.; Du, G.; Lü, K.; Wang, L.; Du, J. A survey and an empirical evaluation of multi-view clustering approaches. *Acm Comput. Surv.* **2024**, *56*, 1–38.
3. Zhang, C.; Chen, H.; Li, H.; Chen, C. Learning latent disentangled embeddings and graphs for multi-view clustering. *Pattern Recognit.* **2024**, *156*, 110839. [\[CrossRef\]](#)
4. Chaudhuri, K.; Kakade, S.M.; Livescu, K.; Sridharan, K. Multi-view clustering via canonical correlation analysis. In Proceedings of the International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 129–136.
5. Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K. Deep canonical correlation analysis. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1247–1255.
6. Kumar, A.; Rai, P.; Daume, H. Co-regularized multi-view spectral clustering. In Proceedings of the International Conference on Neural Information Processing Systems, Granada, Spain, 12–15 December 2011; pp. 1413–1421.
7. Yang, T.; Wang, C.; Guo, J.; Li, X.; Chen, M.; Dang, S.; Chen, H. Triplets-based large-scale multi-view spectral clustering. *Inf. Fusion* **2025**, *121*, 103134.
8. Dou, Z.; Peng, N.; Hou, W.; Xie, X.; Ma, X. Learning multi-level topology representation for multi-view clustering with deep non-negative matrix factorization. *Neural Netw.* **2025**, *182*, 106856. [\[PubMed\]](#)
9. Che, H.; Li, C.; Leung, M.-F.; Ouyang, D.; Dai, X.; Wen, S. Robust hypergraph regularized deep non-negative matrix factorization for multi-view clustering. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *9*, 1817–1829.
10. Chen, R.; Tang, Y.; Zhang, W.; Feng, W. Adaptive-weighted deep multi-view clustering with uniform scale representation. *Neural Netw.* **2024**, *171*, 114–126. [\[PubMed\]](#)
11. Hu, D.; Dong, Z.; Liang, K.; Yu, H.; Wang, S.; Liu, X. High-order topology for deep single-cell multi-view fuzzy clustering. *IEEE Trans. Fuzzy Syst.* **2024**, *32*, 4448–4459.
12. Zhao, H.; Ding, Z.; Fu, Y. Multi-view clustering via deep matrix factorization. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 2921–2927.
13. Yin, Q.; Wu, S.; Wang, L. Multiview clustering via unified and view-specific embeddings learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 5541–5553.
14. Wang, X.; Li, H.; Cao, Y.; Yin, Q. VDMN: View-driven modular network for consistency and complementarity mining in multi-view clustering. In Proceedings of the International Conference on Computer and Communication Systems, Chengdu, China, 18–21 April 2025; pp. 414–419.
15. Chen, J.; Huang, A.; Gao, W.; Niu, Y.; Zhao, T. Joint shared-and-specific information for deep multi-view clustering. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 7224–7235.
16. Zhao, M.; Yang, W.; Nie, F. Deep multi-view spectral clustering via ensemble. *Pattern Recognit.* **2023**, *144*, 109836. [\[CrossRef\]](#)
17. Wan, X.; Liu, J.; Gan, X.; Liu, X.; Wang, S.; Wen, Y.; Wan, T.; Zhu, E. One-step multi-view clustering with diverse representation. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 5774–5786. [\[CrossRef\]](#)
18. Luo, S.; Zhang, C.; Zhang, W.; Cao, X. Consistent and specific multi-view subspace clustering. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 3730–3737.

19. Liang, Y.; Huang, D.; Wang, C.-D. Consistency meets inconsistency: A unified graph learning framework for multi-view clustering. In Proceedings of the IEEE International Conference on Data Mining, Beijing, China, 8–11 November 2019.
20. Kang, Z.; Shi, G.; Huang, S.; Chen, W.; Pu, X.; Zhou, J.T.; Xu, Z. Multi-graph fusion for multi-view spectral clustering. *Knowl.-Based Syst.* **2020**, *189*, 105102. [[CrossRef](#)]
21. Xu, K.; Tang, K.; Su, Z.; Tan, H. Clean and robust multi-level subspace representations learning for deep multi-view subspace clustering. *Expert Syst. Appl.* **2024**, *252*, 124243. [[CrossRef](#)]
22. Kang, Z.; Shi, G.; Huang, S.; Chen, W.; Pu, X.; Zhou, J.T.; Xu, Z. Cluster-graph convolution networks for robust multi-view clustering. *Knowl.-Based Syst.* **2025**, *327*, 114163.
23. Liu, Y.; Chen, J.; Lu, Y.; Ou, W.; Long, Z.; Zhu, C. Adaptively topological tensor network for multi-view subspace clustering. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 5562–5575. [[CrossRef](#)]
24. Kong, Z.; Fu, Z.; Chang, D.; Wang, Y.; Zhao, Y. One for all: A novel dual-space co-training baseline for large-scale multi-view clustering. *arXiv* **2024**, arXiv:2401.15691.
25. Kumar, A.; Daume, H. A co-training approach for multi-view spectral clustering. In Proceedings of the International Conference on Machine Learning, Bellevue, DC, USA, 28 June–2 July 2011; pp. 393–400.
26. Liu, T.-L. Guided co-training for large-scale multi-view spectral clustering. *arXiv* **2017**, arXiv:1707.09866v1.
27. Yang, G.; Zou, J.; Chen, Y.; Du, L.; Zhou, P. Heat kernel diffusion for enhanced late fusion multi-view clustering. *IEEE Signal Process. Lett.* **2024**, *31*, 2310–2314. [[CrossRef](#)]
28. Gan, Y.; You, Y.; Huang, J.; Xiang, S.; Tang, C.; Hu, W.; An, S. Multi-View Clustering via Multi-Stage Fusion. *IEEE Trans. Multimed.* **2025**, *27*, 4571–4583. [[CrossRef](#)]
29. Bruno, E.; Marchand-Maillet, S. Multiview clustering: A late fusion approach using latent models. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA, 19–23 July 2009; pp. 19–23.
30. Liu, X.; Zhu, X.; Li, M.; Wang, L.; Tang, C.; Yin, J.; Shen, D.; Wang, H.; Gao, W. Late fusion incomplete multi-view clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2410–2423. [[CrossRef](#)]
31. Chao, G.; Jiang, Y.; Chu, D. Incomplete contrastive multi-view clustering with high-confidence guiding. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024.
32. Wang, H.; Wang, Q.; Miao, Q.; Ma, X. Joint learning of data recovering and graph contrastive denoising for incomplete multi-view clustering. *Inf. Fusion* **2024**, *104*, 102155. [[CrossRef](#)]
33. Zhu, P.; Yao, X.; Wang, Y.; Hui, B.; Du, D.; Hu, Q. Multi-view deep subspace clustering networks. *IEEE Trans. Cybern.* **2024**, *54*, 4280–4293. [[CrossRef](#)] [[PubMed](#)]
34. Li, Z.; Wang, Q.; Tao, Z.; Gao, Q.; Yang, Z. Deep adversarial multi-view clustering network. In Proceedings of the International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 2952–2958.
35. Mu, S.; Lin, S. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv* **2025**, arXiv:2503.07137.
36. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv* **2017**, arXiv:1701.06538.
37. Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; Huang, J. A survey on mixture of experts in large language models. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 3896–3915. [[CrossRef](#)]
38. Zhang, Y.; Cai, J.; Wu, Z.; Wang, P.; Ng, S.-K. Mixture of experts as representation learner for deep multi-view clustering. In Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; pp. 22704–22713.
39. Feng, J.; Yang, Y.; Xie, Y.; Li, Y.; Guo, Y.; Guo, Y.; He, Y.; Xiang, L.; Ding, G. Debaised novel category discovering and localization. In Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 1753–1760.
40. Wang, G.; Wang, K.; Wang, G.; Torr, P.H.; Lin, L. Solving Inefficiency of Self-supervised Representation Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 9505–9515.
41. Zhan, K.; Shi, J.; Wang, J.; Wang, H.; Xie, Y. Adaptive structure concept factorization for multiview clustering. *Neural Comput.* **2018**, *30*, 1080–1103. [[CrossRef](#)]
42. Cui, C.; Ren, Y.; Pu, J.; Li, J.; Pu, X.; Wu, T.; Shi, Y.; He, L. A novel approach for effective multi-view clustering with information-theoretic perspective. *Neural Inf. Process. Syst.* **2023**, *36*, 44847–44859.
43. Xu, J.; Li, C.; Peng, L.; Ren, Y.; Shi, X.; Shen, H.T.; Zhu, X. Adaptive feature projection with distribution alignment for deep incomplete multi-view clustering. *IEEE Trans. Image Process.* **2023**, *32*, 1354–1366. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.