EASEMVC:Efficient Dual Selection Mechanism for Deep Multi-View Clustering

Baili Xiao[1], Zhibin Dong[1],Ke Liang[1], Suyuan Liu[1], Siwei Wang[2,*], Tianrui Liu[1,*], Xingchen Hu[1]
En Zhu[1,*], Xinwang Liu[1]
[1]National University of Defense Technology, Changsha, China
[2]Intelligent Game and Decision Lab, Beijing, China
{xiaobaili,trliu,wangsiwei13,xinwangliu,enzhu}@nudt.edu.cn

Abstract

Multi-view clustering (MVC) has emerged as a leading paradigm in unsupervised learning, gaining significant attention. Central to this framework is the concept of view-pair contrastive learning, which aims to maximize mutual information between pairs of views, thereby facilitating consistent latent representations. Nevertheless, two critical challenges remain: i) Identifying the most suitable pairs of views for contrastive learning becomes difficult when more than two views are available, especially in the absence of prior knowledge; ii)Including all available views in contrastive learning can degrade performance due to the presence of low-quality views. To address these issues, we propose a novel mechanism, ***EASEMVC****(****E****fficient Du****A****l* ***S****election M****E****chanism for Deep* ***M****ulti-****V****iew* ***C****lustering). EASEMVC begins by constructing a view graph using the Optimal Transport (OT) distance between bipartite graphs of individual views. A view selection module is then designed to perform efficient view-level selection based on the topological relationships within the view graph. Additionally, a cross-view sample graph is built at the sample level, where the topological relationships among samples are used to generate reliable learning weights. Leveraging the selected view pairs and sample weights, contrastive learning is employed to obtain consistent representations across views. Extensive experiments across six benchmark datasets demonstrate that EASEMVC outperforms current state-of-the-art methods.*

1. Introduction

The proliferation of network media and sensor technologies has made multi-view data readily accessible in the information age [9, 18, 57]. Such data encompasses diverse modalities, including textual categories in the 20 Newsgroups dataset [16] and visual features such as HOG [6], SIFT [7], and LBP [25] in Caltech101 [17]. Single-view representations inherently lack completeness due to their limited perspective, restricting their ability to capture comprehensive data characteristics. Multi-view learning addresses this limitation by aggregating complementary information from heterogeneous sources, thereby enhancing downstream tasks like clustering through enriched semantic diversity and cross-view consistency.

Despite the potential of multi-view clustering (MVC)[8, 9, 35, 36, 38–40, 48, 50, 57], existing approaches face a critical challenge: how to effectively reconcile view disparities while preserving view-specific information during cross-view consistency learning. Current methods often indiscriminately fuse all views, disregarding quality variations and redundancy, which degrades clustering performance as noisy or redundant views corrupt shared representations.

Modern deep MVC frameworks typically integrate five components: view-specific feature extractors, single-view self-supervised learning modules(SV-SSL), multi-view self-supervised learning modules(MV-SSL), fusion modules and clustering modules [36]. MV-SSL dominates current research, operating within an information-theoretic framework [13, 30, 31, 37] and manifesting in three paradigms:Contrastive learning-based methods [2, 12, 29] align views by minimizing feature discrepancies.Mutual information maximization methods [8, 19, 20] optimize inter-view dependencies.Information bottleneck-based methods [10, 41, 50] balance compression and relevance.While these 'view-pair' MVC methods excel in pairwise settings, their extension to multi-view scenarios remains contentious. For instance, DCP [20] propose 'complete graph' and 'core view' strategies, yet these ignore prerequisites for valid view-pair learning, risking information loss when handling non-redundant views.

A critical limitation persists: current methods assume uniform view quality and inherent redundancy, leading to two unresolved issues: i)Degradation of high-quality views: Forced alignment of disparate views (e.g., $\mathbf{V}_4$ and $\mathbf{V}_2$ in Figure 1) distorts view-specific information.ii)Sensitivity

*Corresponding author

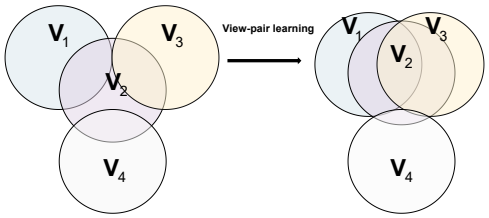


Figure 1. Original view representation and view-pair learning representation. Through view-pair contrastive learning, specific information of the view is disrupted.

to noisy/redundant views: Arbitrary view pairing (e.g., random core view selection) amplifies error propagation.Contrary to the prevailing assumption that increasing views monotonically improves clustering, empirical evidence [35, 36] demonstrates performance degradation in non-redundant multi-view scenarios. This discrepancy motivates our hypothesis: view-pair MVC efficacy depends on multi-view redundancyit enhances performance when redundancy exists but harms it otherwise.

To verify the aforementioned hypothesis, we conducted a preliminary experiment on the Caltech-5V dataset. Integrating essential components from DeepMVC[36], we developed a CoMVC-like[35] framework. The performance of individual views was systematically evaluated before and after view-pair MVC implementation using three standard metrics: Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). As illustrated in Figure 2, the experimental results demonstrate that following multi-view contrastive learning, the quality of View 1 suffered significant degradation. Concurrently, the clustering accuracy of multi-view fused features exhibited a notable decrease from 84.57% to 73.79%.

Therefore, we revisit the essence of multi-view contrastive learning paradigms from an information-theoretic perspective. Our analysis reveals that existing paradigms universally rely on the multi-view redundancy assumption. However, this assumption inherently exhibits a non-transitive property, which serves as the fundamental cause of representation degradation in current methods due to their excessive pursuit of view consistency. Comprehensive theoretical derivations and proofs supporting this conclusion are provided in the supplementary.

Building upon our theoretical insights, we propose an efficient dual selection mechanism that dynamically selects appropriate view pairs for cross-view learning without prior knowledge. This mechanism mitigates the negative impact of low-quality views on high-quality views, and adjusts weights based on the optimal transport distance between view pairs to constrain mutual information during learning. Specifically, we generate anchors for each view and construct the corresponding bipartite graphs. The optimal transport distance is used to quantify the dissimilarity between the bipartite graphs, from which we build a view-graph topology. This structure is then employed to select the most comparable view pairs for further learning. Meanwhile, we design a unified representation learning framework that integrates single-view and cross-view self-supervised learning. The single-view self-supervised learning captures view-specific information, which is then mapped to a shared feature space using a shared MLP. Cross-view sample topologies are built, and view consistency is enforced through weighted sample-level consistency learning. Additionally, clustering-level consistency comparison is incorporated to ensure effective clustering.These components are jointly optimized in an end-to-end manner for mutual benefit. We refer to this framework as **E**fficient du**A**l **S**election m**E**chanism for deep **M**ulti-**V**iew **C**lustering(EASEMVC). Extensive experiments on several challenging multi-view datasets demonstrate that EASEMVC significantly outperforms numerous state-of-the-art traditional and deep MVC methods. The main contributions of this work are as follows:

- We propose an efficient view selection mechanism that facilitates the selection of cross-view pairs for consistency learning without relying on prior knowledge. This module can be seamlessly integrated into existing multi-view clustering (MVC) frameworks based on view-pairs, enabling enhanced cross-view learning.
- We introduce a dynamic weight allocation strategy that assigns adaptive weights to positive sample pairs according to their topological relationships and connectivity within the graph structure. This approach allows the model to adjust the contribution of each sample pair during cross-view contrastive learning, thereby improving the robustness and accuracy of the learning process.
- We present an end-to-end deep multi-view learning framework that jointly learns low-level features, high-level semantics, view-specific information, and clustering, thereby enabling mutual enhancement between these components. Extensive experiments on multiple benchmark multi-view datasets demonstrate that our proposed method significantly outperforms several state-of-the-art (SOTA) traditional and deep MVC methods.

2. Related Work

The rapid advancement of Internet technologies and data acquisition methods has significantly increased the prevalence of multi-view data. Multi-view clustering (MVC) [9, 24, 32, 43, 51, 55, 58] has emerged as a powerful unsupervised learning paradigm that leverages complementary and

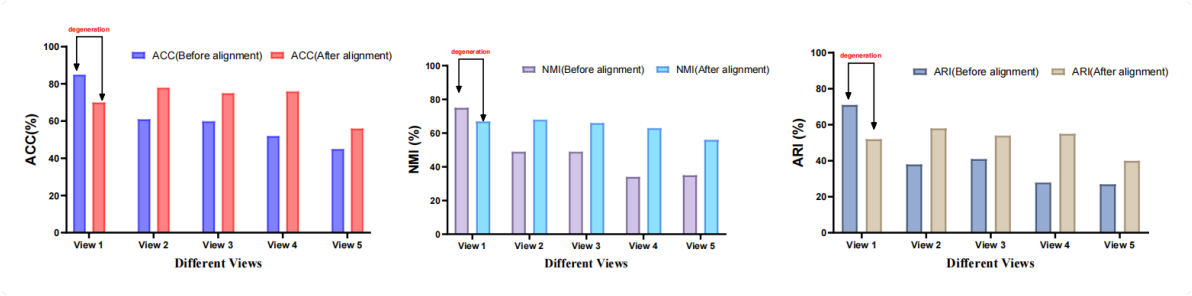


Figure 2. As the high-quality view is constrained to align with low-quality counterparts during the alignment process, View 1 exhibits marked quality degradation.

consistent information across multiple data perspectives.

The remarkable success of deep learning, particularly in unsupervised representation learning [3, 4, 28], has driven substantial progress in deep multi-view clustering. These approaches [5, 22, 23, 27, 42] employ deep neural networks to extract view-specific features while optimizing joint representations to enhance clustering performance.

Contrastive learning represents successful unsupervised learning paradigm [11, 26, 34, 45, 52], which has also been applied to multi-view clustering. The fundamental concept can be traced back to [1], where it was proposed that, in the context of unsupervised learning, invariant representations can be learned by comparing independent but related data points. This can be considered to be equivalent to the process of maximizing the mutual information between different views of the same scene. The classic loss function of contrastive learning, InfoNCE, is essentially designed to maximize the lower bound of mutual information. MVC methods utilizing contrastive learning categorize samples from the same view as positive pairs, whereas the remaining sample pairs (comprising different samples within and across views) are designated as negative pairs. [2, 12, 29] calculate the contrastive loss of different view representations directly, with the objective of aligning view-specific feature representations. [8, 19, 20] directly seek to maximize the mutual information between different views. The information bottleneck method is based on a fundamentally different principle than contrastive learning. The information bottleneck method seeks to identify a compact representation T that maximizes the compression of the source variable X into T while retaining as much relevant information as possible about another variable Y. The original information bottleneck method was developed in the context of supervised learning, where the source variable X serves as data samples and the other variable Y as labels.[10] initially developed the information bottleneck method for application in unsupervised multi-view learning. As previously outlined in Section 1, these methodologies can be collectively termed view-pair self-supervised MVC. The loss function calculation processes for these three methods are detailed in the supplementary materials.

Notably, existing methods lack a unified framework that simultaneously models view-level and sample-level topological structures through bipartite graph representations. Our proposed information-theoretic framework addresses this gap by integrating selective information extraction with consistency optimization, effectively eliminating redundant information while enhancing consensus representation robustness. This synergistic approach enables mutual reinforcement between structural preservation and discriminative feature learning.

3. Method

This section first describes the problem, Then, we introduce our proposed EASEMVC framework and elaborate on each module of the proposed network framework, including a view selection module, a sample weight learning module, a contrastive Learning module, and a clustering module.The EASEMVC framework is illustrated in Figure 3.

3.1. Problem Statement

Suppose the dataset $\mathbf{X} = \left\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(V)}\right\}$ consists of V views, where for the data samples in the v-th view, $\mathbf{X}^{(v)} = \left\{\mathbf{x}_1^{(v)}, \mathbf{x}_2^{(v)}, \ldots, \mathbf{x}_n^{(v)}\right\} \in \mathbb{R}^{n \times d^v}$. Here, n denotes the number of data samples in each view, and d^v represents the feature dimensionality of the v-th view. We assume that the dataset contains K ground-truth clusters.

3.2. Framework Overview

If the assumption of multi-view mutual redundancy does not hold, issues such as degraded contrast learning and misalignment of representation distributions may arise, complicating the clustering process. To address this challenge, we propose a novel framework, as illustrated in Figure 3. This framework consists of a view selection module, a sample weight learning module, a contrastive learning module, and a clustering module. The clustering module enables dual selective fusion of consistent information based on both view

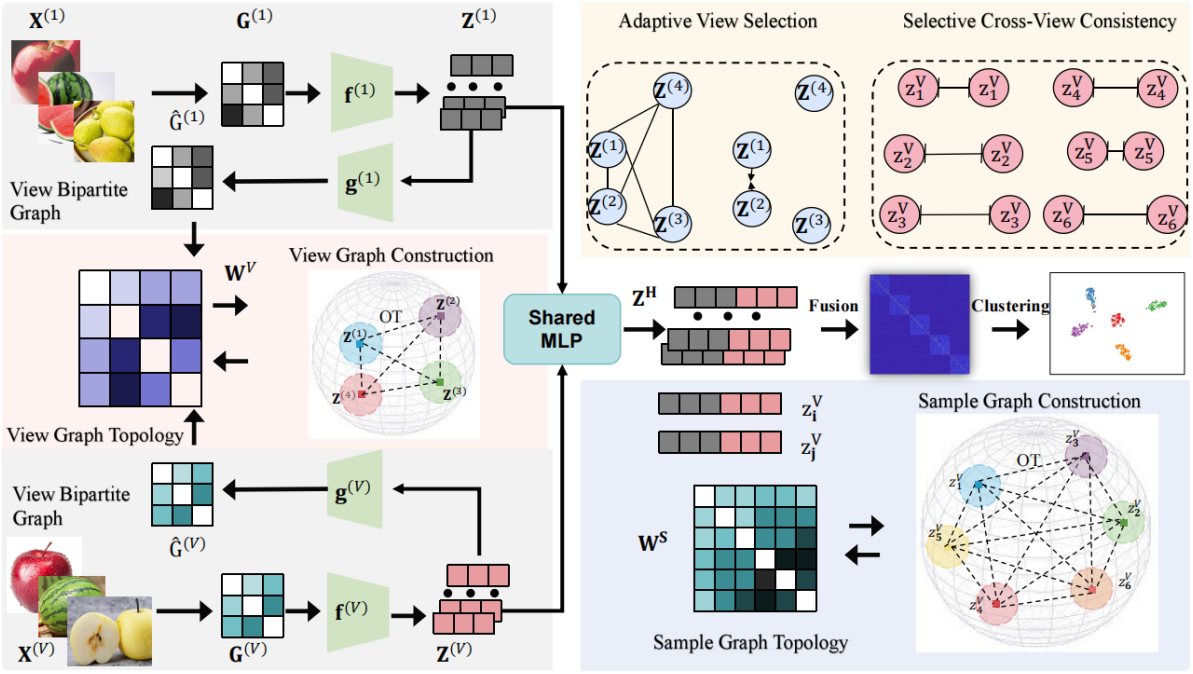


Figure 3. Our EASEMVC framework. The dual-selective contrastive learning mechanism is reflected in the adaptive view selection based on the view topological graph and the cross-view selection consistency based on the sample topological graph

and sample topologies, thereby improving the quality of multi-view clustering.

3.3. Adaptive View Selection on Graph Topology

Inspired by [8, 54], we propose a method for obtaining feature representations for each view through view-independent bipartite graph learning. Specifically, we define a set $\mathbf{G} = \{\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \ldots, \mathbf{G}^{(V)}\}$, where $\mathbf{G}^{(i)}$ represents the bipartite graph for the i-th view.

To effectively construct each bipartite graph, we first employ a clustering algorithm to identify key anchor points within each view. These anchor points serve as foundational nodes, aiding in the capture of structural information within the views. Subsequently, we initialize the bipartite graph for the i-th view as $\mathbf{G}^{(i)} \in \mathbb{R}^{n \times k^{(i)}}$, where n denotes the number of samples and $k^{(i)}$ represents the number of anchor points in the i-th view.

Next, to enhance the accuracy and completeness of the feature representations, we utilize an encoder-decoder architecture to perform mapping and reconstruction within their respective feature spaces for each view. During the encoding phase, the input data is mapped to a latent space to extract a low-dimensional representation of high-dimensional features; in the decoding phase, the system employs these low-dimensional representations to reconstruct the original data. This iterative process of mapping and reconstruction allows us to obtain refined bipartite graph representations $\hat{\mathbf{G}}^{(i)} = g(f(\mathbf{G}^{(i)}))$, where $g(.)$ and $f(.)$ are decoder and encoder, respectively.

By iteratively repeating this process during the pre-training phase, we continuously update the bipartite graphs for each view, ensuring that the model effectively learns the deep associations and feature integrations within the internal structures of the various views.

$$\mathcal{L}_{REC} = \sum_{i=1}^{V} ||\hat{\mathbf{G}}^{(i)} \hat{\mathbf{G}^{(i)}}^{T} - \mathbf{G}^{(i)} \mathbf{G}^{(i)^{T}}||_F^2 \qquad (1)$$

We construct a view-graph topology by evaluating the bipartite graph transfer distances between various views. This topology effectively captures the distances and similarities between views, providing a comprehensive framework for understanding their interrelationships. For each view, we identify the nearest view based on the calculated distances, selecting it as the optimal contrasting view for enhanced contrastive learning.

By focusing on the nearest view, we ensure that the selected contrasting view shares significant structural similarities with the target view, thereby facilitating more effective feature differentiation and representation learning. This approach leverages the inherent similarities to highlight subtle differences, enhancing the model's ability to learn discriminative features. The view-graph topology thus serves as an effective tool in contrastive learning, enabling the selection of the most appropriate contrasting views to maximize learning efficiency and improve overall model per-

formance. Specifically, we compute the view graph matrix $\mathbf{W}^V \in \mathbb{R}^{V\times V}$:

$$w_{ij}^V = D_{OT}(\hat{\mathbf{G}}^{(i)}, \hat{\mathbf{G}}^{(j)}) \tag{2}$$

Inspired by [8],D_{OT} is the process to calculate the bipartite graph transition distances between $\hat{\mathbf{G}}^{(i)}$ and $\hat{\mathbf{G}}^{(i)}$.For any view i, we can obtain its row vector $\mathbf{W}_{i,:}^V$ from $\mathbf{W}^V$ and then select the nearest view j based on distance. That is, selecting the one with the minimum OT distance.

Ultimately, the View Similarity Measure (VSM) outputs a vector $\mathbf{n}$ of length V, where each element $n^{(i)}$ represents the index of the view that is most similar to view i,*i.e.*,$\mathbf{n} = [n^{(1)}, n^{(2)}, \ldots, n^{(V)}]$.

3.4. Selective Cross-View Consistency On Sample Graph Topology

We input the view-specific features $\{\mathbf{Z}^{(v)}\}_{v=1}^V$ obtained during the pre-training phase into a shared Multi-Layer Perceptron (MLP). This approach aims to alleviate the potential conflict between the consistency objective and the reconstruction objective. By utilizing a shared MLP, we can effectively filter out the view-specific information from $(\{\mathbf{Z}^{(v)}\}_{v=1}^V$, allowing for cross-view analysis of samples in a common high-dimensional feature space. The advantage of using a shared MLP lies in its ability to project features from different views into a unified feature space, ensuring that the features across views are comparable.

In this way, we obtain the cross-view high-level features of the samples, denoted as $\mathbf{Z}^H \in \mathbb{R}^{N\times V\times D}$, where N represents the number of samples, V denotes the number of views, and D is the dimensionality of the output features from the shared MLP.

We continuously update the cross-view high-level features of the samples, denoted as $\mathbf{Z}^H \in \mathbb{R}^{N\times V\times D}$, $\mathbf{Z}^H = \{\mathbf{z}_1^H, \mathbf{z}_2^H, ...\mathbf{z}_N^H\}$ for each sample i, the cross-view feature $\mathbf{z}_i^H$ using the representation learning module to obtain refined representations $\hat{\mathbf{z}}_i^H$. We then utilize the Optimal Transport (OT) distance to measure the relationship between two samples i and j, and obtain the sample graph matrix $\mathbf{W}^S \in \mathbb{R}^{N\times N}$, where $w_{ij}^S = D_{OT}(\hat{\mathbf{z}}_i^H, \hat{\mathbf{z}}_j^H)$.

In contrastive learning, a sample with a smaller distance to others is considered more representative, as it provides more meaningful similarity information. Thus, closer samples are more suitable for contrastive learning as they provide more meaningful similarity information. We measure the weight of the *i*-th sample as:

$$w_i = \frac{\sum_{k\neq l} w_{kl}^S}{\sum_{j\neq i} w_{ij}^S} \tag{3}$$

where, w_i represents the weight of the i-th sample,$\sum_{j\neq i} w_{ij}^S$ is the sum of distances from the i-th sample to other samples, and $\sum_{k\neq l} w_{kl}^S$ is the total sum of distances between all samples. Perform softmax normalization on the weights of each sample:

$$\hat{w}_i = \frac{e^{w_i}}{\sum_{j=1}^N e^{w_j}} \tag{4}$$

3.5. Dual Selection Contrastive Learning Module

In our approach, we first utilize the view selection module to choose the most similar view for each view. This process ensures that in contrastive learning, each view can be compared with its most relevant counterpart, thereby enhancing the model's ability to capture similarity information. Simultaneously, the sample weight learning module assigns appropriate weights to each pair of samples, reflecting the importance of samples in the learning process. This allows us to focus more on those samples that contribute significantly to model training. Subsequently, we employ constrained weighted contrastive learning to extract optimal consistency information between views. This method not only strengthens the consistency among different views but also improves the model's robustness and accuracy when handling complex data.

The multi-view contrastive learning loss function that we designed is as follows:

$$\mathcal{L}_{CL} = \sum_{m=1}^{V}\sum_{n=1}^{V} \mathbf{1}(n = n^{(V)})\ell(\mathbf{z}^m, \mathbf{z}^n), \tag{5}$$

where, $\ell(\mathbf{z}^m, \mathbf{z}^n)$ represents the contrastive loss between the m-th and n-th views. $\mathbf{1}(n = n^{(V)})$ is the indicator function, which takes the value of 1 when n is the most similar view to the m-th view, and 0 otherwise. The sample-weighted feature contrastive loss is formulated as follows:

$$\ell(\mathbf{z}^m, \mathbf{z}^n) = -\frac{1}{N}\sum_{i=1}^{N} log\frac{\hat{w}_i e^{\frac{s(\mathbf{z}_i^m, \mathbf{z}_i^n)}{\tau}}}{\sum_{j=1}^N \sum_{V=m,n} \frac{s(\mathbf{z}_i^m, \mathbf{z}_i^n)}{\tau} - \frac{1}{\tau}} \tag{6}$$

Where τ denotes the temperature parameter, $\hat{w}_i$ denotes softmax normalization of the weights of each sample, $s(\mathbf{z}_i^m, \mathbf{z}_i^n)$ denotes the cosine similarity between sample from two views.

3.6. Clustering Module

We then introduce an effective Deep Divergence Clustering (DDC) loss [14] to implement the clustering module. The introduced DDC loss promotes inter-cluster separability and intra-cluster compactness.

Here is a detailed description of these three components along with their specific formulas:

1. Information-Theoretic Divergence Loss: This component aims to promote cluster separation by minimizing the divergence between different clusters. Common divergence measures include Kullback-Leibler (KL) divergence

or Jensen-Shannon divergence. Assuming P and Q are the probability distributions of two clusters, the divergence loss can be expressed as:

$$L_{\text{divergence}} = D_{\text{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (7)$$

2. Intra-Cluster Compactness Loss: This component aims to minimize the distance between data points within the same cluster to ensure compactness. This can be achieved by minimizing the Euclidean distance between data points within a cluster:

$$L_{\text{compactness}} = \sum_k \sum_{v_i, v_j \in C_k} \|v_i - v_j\|^2 \quad (8)$$

where C_k represents the k-th cluster, and $\|v_i - v_j\|$ denotes the Euclidean distance between data points v_i and v_j.

3. Geometric Structure Utilization Loss: This component leverages the geometric structure of the clustering space by considering the spatial relationships of data points in the feature space. It can be achieved by maintaining the relative distances between data points:

$$L_{\text{geometry}} = \sum_i \sum_j \left(\frac{\|v_i - v_j\|}{d_{\max}} \right) \cdot \mathbb{1}(v_i \in C_k, v_j \in C_l) \quad (9)$$

where $d_{\max}$ is the maximum distance between all data points in the feature space, and $\mathbb{1}(v_i \in C_k, v_j \in C_l)$ is the indicator function that takes the value of 1 when v_i and v_j belong to clusters C_k and C_l, respectively, and 0 otherwise.

By combining these three components, the DDC loss function can simultaneously promote inter-cluster separation and intra-cluster compactness and leverage the geometric structure of the data during the optimization process, thereby enhancing clustering effectiveness. Overall, the DDC loss is given by:

$$\mathcal{L}_{DDC} = \ell_{\text{divergence}} + \ell_{\text{compactness}} + \ell_{\text{geometry}} \quad (10)$$

3.7. Optimization

In summary, the final loss of our EASEMVC is:

$$\mathcal{L} = \mathcal{L}_{REC} + \alpha \mathcal{L}_{CL} + \beta \mathcal{L}_{DDC}, \quad (11)$$

where α and β are hyperparameters that control the contributions of the contrastive learning loss and the DDC clustering loss, respectively. To make a clear presentation,the specific algorithm flow is shown in Algorithm 1.

Algorithm 1 The Proposed EASEMVC

Require: $\mathbf{X}^{(v)} = \left\{ \mathbf{x}_1^{(v)}, \mathbf{x}_2^{(v)}, \ldots, \mathbf{x}_n^{(v)} \right\} \in \mathbb{R}^{n \times d^v}$, number of clusters k

while not reaching the pretrain epochs **do**
  for $i = 1 : V$ **do**
    Obtain the initial bipartite graph $\mathbf{G}^{(i)}$ on the i-th view via k-means.
    The refined bipartite graph $\hat{\mathbf{G}}^{(i)}$ is obtained through Eq.(1).
    The optimal transport (OT) distance w_{ij}^V between the two bipartite graphs $\hat{\mathbf{G}}^{(i)}$ and $\hat{\mathbf{G}}^{(j)}$ is calculated through Eq.(2).
  end for
end while
while not reaching the maximal epochs **do**
  Obtain the sample graph matrix $\mathbf{W}^S$, calculate softmax normalization weights $\hat{w}_i$ of each sample-pair through Eq.(3) and Eq.(4).
  Compute the view reconstruction loss by Eq.(1).
  Compute the dual-selective contrastive learning loss by Eq.(5).
  Compute the DDC clustering loss by Eq.(10).
  Jointly optimize the overall loss Eq.(11) by Adam optimizer.
end while
return Obtaining the final clustering result.

4. Experiments

4.1. Datasets and Settings

In this part of the study, we assess the effectiveness of our introduced model by comparing it to nine other algorithms across six multi-view datasets. Table 2 presents the number of samples(N), number of clusters(C), number of views(V), and feature dimensionality(D) for each dataset.

We implement our EASEMVC model in Pytorch 1.12 and carry all experiments on a Ubuntu-22.04 OS environment with an NVIDIA GeForce RTX 4090 GPU. For model training, we employ the Adam optimizer with its default settings and establish an initial learning rate of 0.0001. Additionally, we configure the batch size to 256. The pretrain phase epoch is 200 and main train epoch is 50. The range for hyper-parameters α and β is [0.001, 1000].

4.2. Comparisons with State of the Arts

For a thorough analysis, we employ four widely-used clustering metrics: accuracy (ACC), normalized mutual information (NMI), purity, and F-score. These metrics are crucial for evaluating clustering performance, with higher values indicating superior results. We conduct a comparative analysis with two state-of-the-art traditional MVC methods and eight deep MVC methods. As shown in Table 1, the performance of our method on all metrics across the five image datasets surpasses other comparative methods, especially on HandWritten, UCI, and ALOI-100. We observe that our method achieved significant success on the ALOI-

Table 1. ACC, NMI, Purity and F-score comparison of different clustering algorithms on benchmark datasets.(The best result is bolded; the second best is underlined) O/M refers to out-of-memory failure. - indicates the error of the method itself.

Datasets	Conventional Methods		Deep Methods								
	LMVSC[15]	BMVC[56]	AE2-Nets[53]	SDMVC[47]	MFLVC[48]	DSMVC[33]	DEMVC[46]	SDSNE[21]	SEM[49]	SCMVC[44]	**EASEMVC**
ACC(%)											
Caltech101-20	47.33	52.64	34.30	49.04	50.92	37.43	33.99	54.19	51.86	47.61	**55.87**
HandWritten	85.10	75.50	87.82	49.45	86.20	95.25	27.45	87.60	95.10	94.85	**96.05**
UCI	89.35	85.45	80.57	63.00	92.00	85.45	62.30	84.55	91.20	92.95	**93.25**
ALOI-100	42.40	63.54	26.90	28.72	23.12	14.89	55.63	O/M	77.78	82.52	**83.81**
BBC	52.41	75.04	29.34	62.48	49.34	42.77	55.62	**85.84**	65.26	74.31	83.94
FashionMNIST	O/M	60.62	59.91	58.86	O/M	58.33	51.28	O/M	61.62	59.85	**62.15**
NMI(%)											
Caltech101-20	48.40	48.92	42.76	64.13	59.35	47.05	7.16	63.03	50.50	62.93	**64.15**
HandWritten	80.56	77.52	80.18	51.9	85.53	91.19	29.39	87.75	88.33	89.63	**92.67**
UCI	83.27	82.52	59.51	64.29	85.40	80.67	60.54	89.05	84.26	87.01	**89.42**
ALOI-100	55.82	76.99	49.47	64.58	67.88	40.57	77.46	O/M	88.71	92.36	**93.17**
BBC	43.00	49.26	1.73	54.96	43.14	13.96	32.20	**73.40**	57.26	53.6	66.58
FashionMNIST	O/M	**65.28**	60.66	64.42	O/M	58.33	59.74	O/M	61.95	58.73	64.50
Purity(%)											
Caltech101-20	67.22	62.07	63.97	76.99	70.75	67.10	33.99	73.09	64.15	76.7	**77.98**
HandWritten	85.10	79.45	87.82	49.85	86.20	95.25	27.45	87.60	87.15	94.85	**96.65**
UCI	89.35	85.45	80.57	67.05	92.00	85.40	63.15	87.05	91.20	92.95	**93.30**
ALOI-100	44.09	65.49	27.58	31.09	23.12	15.79	66.86	O/M	80.00	84.36	**85.48**
BBC	56.20	75.04	35.69	75.47	64.46	47.59	57.23	**85.84**	77.08	74.31	83.94
FashionMNIST	O/M	**67.30**	63.27	67.05	O/M	56.86	63.37	O/M	63.41	61.60	66.18
F-score(%)											
Caltech101-20	41.86	**65.70**	34.20	55.78	54.13	42.29	35.86	59.70	63.20	57.42	60.36
HandWritten	77.23	71.13	79.35	42.10	85.82	95.27	17.66	87.83	82.09	89.48	**96.65**
UCI	81.28	78.72	67.65	62.77	92.05	80.67	60.54	89.05	88.62	89.13	**89.42**
ALOI-100	31.68	50.98	24.59	O/M	12.02	0.36	66.71	O/M	78.24	80.47	**82.83**
BBC	44.03	62.62	27.03	65.55	43.14	53.32	59.42	**85.38**	75.47	80.20	83.72
FashionMNIST	O/M	54.87	55.35	56.09	O/M	58.03	48.39	O/M	56.83	55.35	**61.50**

Table 2. Multi-view datasets used in experiments.

Dataset	N	C	V	D
Caltech101-20	2386	20	6	48/40/254/1984/512/928
HandWritten	2000	10	6	216/76/64/6/240/47
UCI	2000	10	3	64/76/216
ALOI-100	10800	100	4	77/13/64/125
BBC	685	5	4	4659/4633/4665/4684
FashionMNIST	70000	10	4	944/576/512/640

Table 3. Ablation study on ALOI-100 dataset.

Components			Metrics			
$\mathcal{L}_{rec}$	$\mathcal{L}_{CL}$	$\mathcal{L}_{DDC}$	ACC	NMI	Purity	F-score
✓			70.59	86.45	72.56	67.78
✓	✓		75.79	84.83	72.94	75.14
✓		✓	78.98	90.84	80.96	76.29
✓	✓	✓	**83.81**	**93.17**	**85.48**	**82.83**

100 dataset, improving the accuracy (ACC) by 20 points compared to the second-best method. For text datasets, all metrics are second best, but the gap with the optimal method is minimal.

4.3. Ablation Studies

Representation Learning and Clustering Module: To gain deeper insights into the significance of each module within EASEMVC, we undertake a comprehensive ablation study using the ALOI-100 and BBC datasets. The specific results are delineated in Tables 3 and Tables 4. Our observations reveal that each module can considerably enhance clustering performance.

Further, the following seven experiments are designed to substantiate the efficacy of adaptive view selection on the view topology graph(VSM) and selective cross-view consistency on the sample topology graph(SSM). We conducted experiments by equally enabling view-level contrastive learning among all views, which eliminated the adaptive view selection on the view topology graph. Additionally, by removing the weights of positive sample pairs, we eliminated the selective cross-view consistency on the sample topology graph. The results, as depicted in Figures 4(a) and (b), clearly demonstrate that employing contrastive learning uniformly across all views leads to a notable degradation in clustering performance. This confirms the importance of adaptive view selection in preserving the quality of the learned representations. In contrast, as illustrated in Figures 4(c) and (d), our proposed dual-selection mechanism contrastive learning method effectively mitigates this learning degradation phenomenon. Specifically, both the adaptive view selection on the view topology graph and the selective cross-view consistency on the sample topology graph are shown to be effective, with the former demonstrating a more significant impact.

Parameter Sensitivity Analysis: To evaluate the impact of hyperparameters α and β on the contrastive learning and DDC clustering losses, we varied their values within

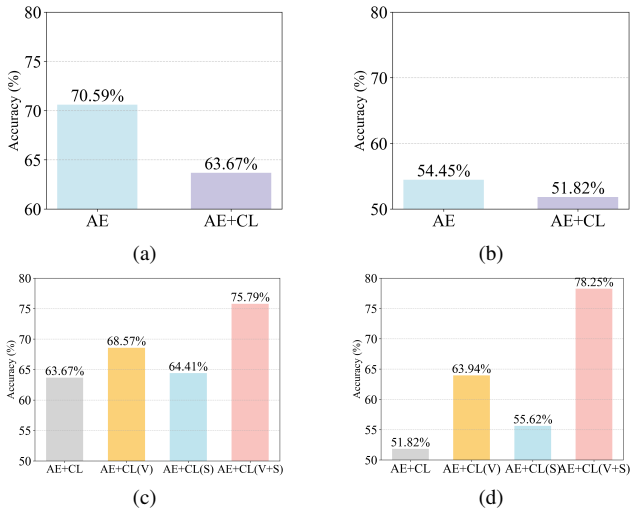
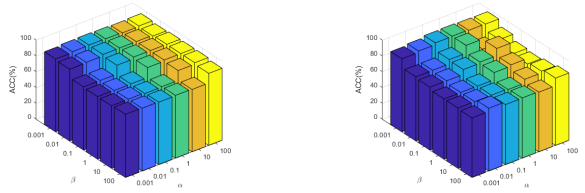


Figure 4. Ablation study of contrastive learning and dual selective mechanism on ALOI-100 and BBC dataset. AE denotes Autoencoder.CL denotes contrastive learning. V denotes adaptive view selection on the view topology graph(VSM).S denotes selective cross-view consistency on the sample topology graph(SSM)

Table 4. Ablation study on BBC dataset.

Components			Metrics			
$\mathcal{L}_{rec}$	$\mathcal{L}_{CL}$	$\mathcal{L}_{DDC}$	ACC	NMI	Purity	F-score
✓			54.45	32.7	54.45	55.17
✓	✓		78.25	61.56	78.25	74.90
✓		✓	57.96	34.84	55.47	54.96
✓	✓	✓	**83.94**	**66.58**	**83.94**	**83.72**



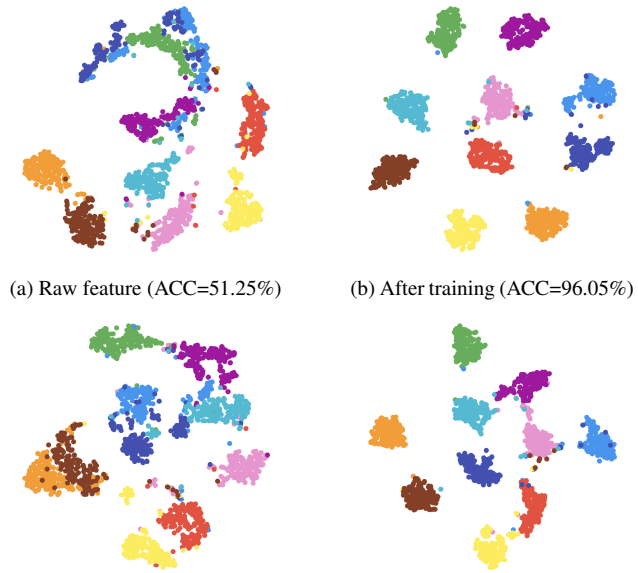
(a) UCI_digit (b) HandWritten

Figure 5. Sensitivity analysis of hyperparameters α and β.

$[0.001, 0.01, 0.1, 1, 10, 100, 1000]$. Figure 5 demonstrates that our method is to the choice of α and β.

4.4. Visualization of Clustering Performance

Furthermore, we visualize the clustering results of EASEMVC on the HandWritten and UCI-digit datasets. This include visualizations of both the original features and the final clusters obtained through our EASEMVC framework, highlighting its effectiveness in improving the clustering performance. As illustrated in Figure 6, it is evident that the learned representations become more compact and discriminative. The final representations successfully cluster similar samples together, resulting in accuracy values of 97.15% and 92.50 % on the HandWritten and UCI-digit datasets, respectively. We have demonstrated that our model can achieve excellent clustering performance.



(a) Raw feature (ACC=51.25%) (b) After training (ACC=96.05%)

(c) Raw feature (ACC=66.85%) (d) After training (ACC=93.25%)

Figure 6. Comparison of features before and after training on HandWritten and UCI-digit dataset

5. Conclusions

To address the issue of learning degradation in view-pair learning when multiview redundancy is not satisfied, we propose a dual selective mechanism contrastive learning multiview clustering framework. This framework is based on adaptive view selection on the view topology graph and selective cross-view consistency on the sample topology graph. Extensive experiments have demonstrated the superiority of our method compared to traditional and deep state-of-the-art (SOTA) methods. We believe our motivation merits discussion and will offer valuable new insights to the multi-view clustering community.

6. Acknowledgment

This work is supported by the National Key R&D Program of China under Grant No.2022ZD0209103, the National Natural Science Foundation of China (project no. 62325604, 62276271, 62406329, 62476280,62476281, 62441618), National Natural Science Foundation of China Joint Found under Grant No. U24A20323.

References

[1] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993. 3

[2] Bowei Chen, Sen Xu, Heyang Xu, Xuesheng Bian, Naixuan Guo, Xiufang Xu, Xiaopeng Hua, and Tian Zhou. Structural deep multi-view clustering with integrated abstraction and detail. *Neural Networks*, 175:106287, 2024. 1, 3

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 3

[5] Chenhang Cui, Yazhou Ren, Jingyu Pu, Jiawei Li, Xiaorong Pu, Tianyi Wu, Yutao Shi, and Lifang He. A novel approach for effective multi-view clustering with information-theoretic perspective. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 886–893. Ieee, 2005. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[8] Zhibin Dong, Siwei Wang, Jiaqi Jin, Xinwang Liu, and En Zhu. Cross-view topology based consistent and complementary information for deep multi-view clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19440–19451, 2023. 1, 3, 4, 5

[9] Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12350–12368, 2023. 1, 2

[10] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020. 1, 3

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[12] Shizhe Hu, Guoliang Zou, Chaoyang Zhang, Zhengzheng Lou, Ruilin Geng, and Yangdong Ye. Joint contrastive triple-learning for deep multi-view clustering. *Information Processing & Management*, 60(3):103284, 2023. 1, 3

[13] Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, and Yangdong Ye. A survey on information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[14] Michael Kampffmeyer, Sigurd Løkse, Filippo M Bianchi, Lorenzo Livi, Arnt-Børre Salberg, and Robert Jenssen. Deep divergence-based approach to clustering. *Neural Networks*, 113:91–101, 2019. 5

[15] Zhao Kang, Wangtao Zhou, Zhitong Zhao, Junming Shao, Meng Han, and Zenglin Xu. Large-scale multi-view subspace clustering in linear time. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4412–4419, 2020. 7

[16] Ken Lang. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier, 1995. 1

[17] Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the AAAI conference on artificial intelligence*, 2015. 1

[18] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883, 2018. 1

[19] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11174–11183, 2021. 1, 3

[20] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4447–4461, 2022. 1, 3

[21] Chenghua Liu, Zhuolin Liao, Yixuan Ma, and Kun Zhan. Stationary diffusion state neural estimation for multiview clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7542–7549, 2022. 7

[22] Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14193–14201, 2024. 3

[23] Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. Simple contrastive multi-view clustering with data-level fusion. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 4697–4705, 2024. 3

[24] Huimin Ma, Siwei Wang, Junpu Zhang, Shengju Yu, Suyuan Liu, Xinwang Liu, and Kunlun He. Symmetric multi-view subspace clustering with automatic neighbor discovery. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[25] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002. 1

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[27] Jingyu Pu, Chenhang Cui, Xinyue Chen, Yazhou Ren, Xiaorong Pu, Zhifeng Hao, S Yu Philip, and Lifang He. Adaptive feature imputation with latent graph for deep incomplete

multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14633–14641, 2024. 3

[28] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, S Yu Philip, and Lifang He. Deep clustering: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 3

[29] Zhenqiu Shu, Bin Li, Cunli Mao, Shengxiang Gao, and Zhengtao Yu. Structure-guided feature and cluster contrastive learning for multi-view clustering. *Neurocomputing*, 582:127555, 2024. 1, 3

[30] Noam Slonim. *The information bottleneck: Theory and applications*. PhD thesis, Citeseer, 2002. 1

[31] Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. In *COLT*, number 114, pages 403–414, 2008. 1

[32] Mengjing Sun, Pei Zhang, Siwei Wang, Sihang Zhou, Wenxuan Tu, Xinwang Liu, En Zhu, and Changjian Wang. Scalable multi-view subspace clustering with unified anchors. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3528–3536, 2021. 2

[33] Huayi Tang and Yong Liu. Deep safe multi-view clustering: Reducing the risk of clustering performance degradation caused by view increase. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 202–211, 2022. 7

[34] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 3

[35] Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1255–1265, 2021. 1, 2

[36] Daniel J Trosten, Sigurd Løkse, Robert Jenssen, and Michael C Kampffmeyer. On the effects of self-supervision and contrastive alignment in deep multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23976–23985, 2023. 1, 2

[37] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020. 1

[38] Xinhang Wan, Jiyuan Liu, Xinbiao Gan, Xinwang Liu, Siwei Wang, Yi Wen, Tianjiao Wan, and En Zhu. One-step multi-view clustering with diverse representation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2024. 1

[39] Xinhang Wan, Jiyuan Liu, Hao Yu, Qian Qu, Ao Li, Xinwang Liu, Ke Liang, Zhibin Dong, and En Zhu. Contrastive continual multiview clustering with filtered structural fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[40] Xinhang Wan, Bin Xiao, Xinwang Liu, Jiyuan Liu, Weixuan Liang, and En Zhu. Fast continual multi-view clustering with incomplete views. *IEEE Transactions on Image Processing*, 33:2995–3008, 2024. 1

[41] Zhibin Wan, Changqing Zhang, Pengfei Zhu, and Qinghua Hu. Multi-view information-bottleneck representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10085–10092, 2021. 1

[42] Jing Wang, Songhe Feng, Gengyu Lyu, and Jiazheng Yuan. Surer: Structure-adaptive unified graph neural network for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15520–15527, 2024. 3

[43] Siwei Wang, Xinwang Liu, Xinzhong Zhu, Pei Zhang, Yi Zhang, Feng Gao, and En Zhu. Fast parameter-free multi-view subspace clustering with consensus anchor guidance. *IEEE Transactions on Image Processing*, 31:556–568, 2021. 2

[44] Song Wu, Yan Zheng, Yazhou Ren*, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, and Lifang He. Self-Weighted Contrastive Fusion for Deep Multi-View Clustering. *IEEE Transactions on Multimedia*, 2024. 7

[45] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3

[46] Jie Xu, Yazhou Ren, Guofeng Li, Lili Pan, Ce Zhu, and Zenglin Xu. Deep embedded multi-view clustering with collaborative training. *Information Sciences*, 573:279–290, 2021. 7

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

[48] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16051–16060, 2022. 1, 7

[49] Jie Xu, Shuo Chen, Yazhou Ren, Xiaoshuang Shi, Hengtao Shen, Gang Niu, and Xiaofeng Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[50] Xiaoqiang Yan, Zhixiang Jin, Fengshou Han, and Yangdong Ye. Differentiable information bottleneck for deterministic multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27435–27444, 2024. 1

[51] Ben Yang, Xuetao Zhang, Feiping Nie, Fei Wang, Weizhong Yu, and Rong Wang. Fast multi-view clustering via nonnegative and orthogonal factorization. *IEEE transactions on image processing*, 30:2575–2586, 2020. 2

[52] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6210–6219, 2019. 3

[53] Changqing Zhang, Yeqing Liu, and Huazhu Fu. Ae2-nets: Autoencoder in autoencoder networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2577–2585, 2019. 7

[54] Chuyu Zhang, Hui Ren, and Xuming He. P $^ 2$ OT: Progressive Partial Optimal Transport for Deep Imbalanced Clustering. *arXiv preprint arXiv:2401.09266*, 2024. 4

[55] Yi Zhang, Fengyu Tian, Chuan Ma, Miaomiao Li, Hengfu Yang, Zhe Liu, En Zhu, and Xinwang Liu. Regularized instance weighting multiview clustering via late fusion alignment. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 2

[56] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1774–1782, 2018. 7

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

[58] Linlin Zong, Xianchao Zhang, and Xinyue Liu. Multi-view clustering on unmapped data via constrained non-negative matrix factorization. *Neural Networks*, 108:155–171, 2018. 2