# Improving multi-view document clustering: leveraging multi-structure processor and hybrid ensemble clustering module

**Ruina Bai**[1,2], **Qi Bai**[3], **Ruizhang Huang**[1,2*] , **Yanping Chen**[1,2], **Yongbin Qin**[1,2*]

1. Text Computing & Cognitive Intelligence Engineering Research Center of National Education Ministry, College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
2. State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China
3. School of Mathematical Sciences, Zhejiang University, Hangzhou, 310058, China

**Abstract**

We introduce a multi-view document clustering model called DMsECN (Deep Multi-structure Ensemble Clustering Network), comprising a multi-structure processor and a hybrid ensemble clustering module. Unlike existing models, DMsECN distinguishes itself by creating a consensus structure from multiple clustering structures. The multi-structure processor comprises two stages, each contributing to the extraction of clustering structures that preserve both consistency and complementarity across multiple views. Representation learning extracts both view and view-fused representations from multi-views through the use of contrastive learning. Subsequently, multi-structure learning employs distinct view clustering guidance to generate the corresponding clustering structures. The hybrid ensemble clustering module merges two ensemble methods to amalgamate multiple structures, producing a consensus structure that guarantees both the separability and compactness of clusters within the clustering results. The attention-based ensemble primarily concentrates on learning the contribution weights of diverse clustering structures, while the similarity-based ensemble employs cluster assignment similarity and cluster classification dissimilarity to guide the refinement of the consensus structure. Experimental results demonstrate that DMsECN outperforms other models, achieving new state-of-the-art results on four multi-view document clustering datasets.

**Keywords:** multi-view document clustering, divergent clustering structure, hybrid ensemble clustering

## 1. Introduction

Multi-view document clustering has gradually become an important task in fields such as text mining and sentiment analysis, owing to the fact that a single document sample can be described from different views (Hassani et al., 2020; Zhang et al., 2022). In addition to the conventional content view, multi-view news documents encompass view delineating their propagation behavior, view elucidating their headlines, view delineating associated news articles, and more. These views depict the news documents from various perspectives that allow us to understand documents comprehensively. The extraction of valuable clustering partitions, by taking into account both consistency and complementarity across all views, has been garnering heightened attention.

Recently, for achieving outstanding clustering performance, various deep multi-view clustering methods have been proposed (Zhu et al., 2019; Bai et al., 2021; Xu et al., 2021a; Yang et al., 2022; Bai et al., 2022; Hu et al., 2023). These approaches amalgamate pre-learned low-dimensional representations from each view to create a view-fused representation that governs cluster partitioning across multiple views. Guided by the consistent clustering objective, both individual view and view-fused representations undergo refinement. Numerous recent studies have demonstrated that enhancing individual view and view-fused representations can improve cluster partitioning accuracy, resulting in superior performance in deep multi-view clustering.

However, acquiring a shared document representation from multiple perspectives necessitates data from various views to maintain an identical manifold structure, a concept that runs counter to the reality that similarity(structure) can fluctuate across different views. Current methodologies cannot ensure the optimality of the consensus structure, as the manifold structure from each view fails to provide mutually support evidence. Adjusting the representation of various views solely based on the final unified clustering objective might result in an excessive reliance on representation consistency, jeopardizing the inherent diversity within multi-view document data and potentially leading to the loss of distinctive features. Hence, determining how to emphasize the divergent manifold structure of each view, all the while preserving consistency in the representation space, constitutes a significant research inquiry within the domain of deep multi-view document clustering tasks.

Unfortunately, the effective utilization of different manifold structure of each document view faces two inescapable issues: (1) How to explore the view-specific clustering structure from various

* Corresponding author



*LREC-COLING 2024*, pages 8866–8876

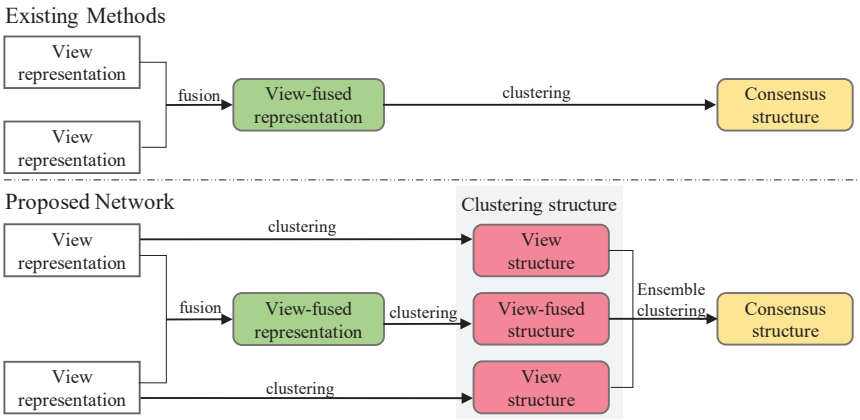



Figure 1: The distinctions between existing methods and the proposed network.

document views? While each view within a particular document sample pertains to the same subject, variations in data type or focal perspective can yield distinct clustering structures among different document views. The initial challenge revolves around the extraction of clustering structures from diverse views, while preserving consistency and complementarity in their features. (2) How to ensemble the multiple clustering structures? A clustering process that emphasizes distinct views independently is bound to result in multiple clustering structures. Hence, the integration of these diverse clustering structures into a coherent clustering arrangement, ultimately yielding the final cluster assignment, holds paramount importance in the comprehensive landscape of multi-view document clustering. As different views exhibit distinct manifold structures, the clustering structure derived from the representation may exhibit inconsistency with the ultimate clustering task objective. Hence, the second issue lies in the formulation of a flexible ensemble aimed at achieving a consensus structure.

To address above issues, we propose a deep multi-structure ensemble clustering network for multi-view document clustering, named DMsECN. The model comprises two components: the multi-structure processor and the hybrid ensemble clustering module. The distinctions between existing methods and the proposed network are visualized in Figure 1. The existing models merge the view representations and employ it directly to generate the consensus structure, preserving solely the inter-view consistency. In contrast, our proposed model acquires a clustering structure for each view representation using multi-structure processor, and ultimately employs a hybrid ensemble clustering module to secure a consensus structure for clustering while simultaneously preserving both consistency and complementarity across different views. Within the multi-structure processor, contrastive learning is employed to bolster representation learning, while distinct view clustering guidance is utilized to generate corresponding clustering structures for all view and view-fused representations. In the hybrid ensemble clustering module, we employ attention-based and similarity-based structure ensemble to produce a consensus structure, which is refined using consensus clustering guidance for clustering.

## 2. Related Work

**Deep Multi-view Clustering** Motivated by the promising progress of deep learning in unsupervised problems, many recent works have been focused on the deep learning-based multi-view clustering. Most of these methods rely on the generative models to learn latent representations from data, such as autoencoder-based methods (Wang et al., 2015; Zhang et al., 2019; Lin et al., 2021; Ke et al., 2022; Abavisani and Patel, 2018), variational autoencoder-based methods (Xu et al., 2021b; Yin et al., 2020), and generative adversarial networks-based methods (Li et al., 2019b; Zhou and Shen, 2020). Inspired by DEC (Xie et al., 2016), a joint framework of deep multi-view clustering is proposed which learns the multiple deep embedded features, multi-view fusion mechanism and clustering assignment simultaneously (Lin et al., 2018). Researchers (Bai et al., 2021) have investigated the extraction of complementary semantic information from high-dimensional sample data space using different enhanced semantic embedders. DEMVC (Xu et al., 2021a) learns feature representation and clustering assignment of all views through collaborative training, to better mine the complementary and consistent information of each view. A novel multi-view clustering method was proposed by learning a shared generative latent representation that obeys a mixture of Gaussian distributions (Yin et al., 2020). In Multi-VAE (Xu et al., 2021b), all views' clus-

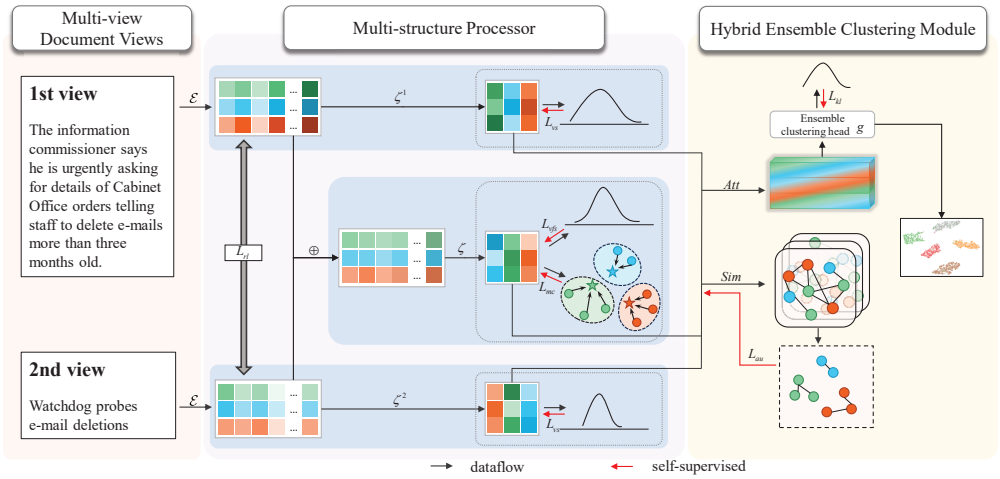



Figure 2: The overall architecture of proposed network.

ter representation and each view's specific visual representations are disentangled by the proposed view-common variable and view-peculiar variables, respectively. In (Li et al., 2019b), generative adversarial networks were utilized to reconstruct the samples from a common representation that shared by multiple views.

**Ensemble Clustering** The purpose of ensemble clustering is to combine multiple base clustering structures into a better and more robust consensus structure (Topchy et al., 2005; Huang et al., 2023). Previous ensemble clustering based methods can mostly be classified into 3 categories: the pair-wise co-occurrence-based methods (Fred and Jain, 2005; Iam-On et al., 2011), the median partition-based methods (Topchy et al., 2005; Huang et al., 2016), and the graph partition-based methods (Strehl and Ghosh, 2002; Huang et al., 2018). And most of them are devised for single-view data. With the emergence of multi-view data, ensemble clustering as an efficient technique to handle multi-view clustering task by utilizing complementary information from multi-view data has gradually gained attention. A multi-view ensemble clustering (MVEC) (Tao et al., 2017) is proposed to learn a consensus clustering from the multiple co-association matrices built in multiple views with low-rank and sparse constraints. Then, Tao et al. (2019) further incorporated marginalized denoising autoencoder into MVEC, and presented a marginalized multi-view ensemble clustering method. Yan et al. (2020) proposed a clustering scheme named synergetic information bottleneck (SIB) for joint multi-view and ensemble clustering. And Niu et al. (2023) proposed a multi-view ensemble clustering approach using joint affinity matrix, which is generated by sample-level weight. In the study (Zhao et al., 2023), a double high-Order correlation preserved robust multi-view ensemble clustering (DC-RMEC) method is devised, which preserves the high-order inter-view correlation and the high-order correlation of original data simultaneously.

## 3. The proposed network

Consider a multi-view document dataset comprising $V$ distinct views, denoted as $X = \{X^1, \cdots, X^V\}$, each containing diverse information. Within each view, there are $N$ samples, represented as $X^v = \{x_i^v\}_{i=1}^N$. The goal of multi-view clustering is to group these multi-view document samples into $K$ clusters by mining various clustering structures. For the sake of simplicity, we employ a dataset with only 2 views as an example.

### 3.1. Overview of DMsECN

The overall structure of DMsECN is shown in Figure 2, including a multi-structure processor and a hybrid ensemble clustering module. The multi-structure processor is comprised of representation learning and multiple structure learning processes. To be more precise, within the representation learning phase, a shared semantic encoder is employed for the extraction of representations from the various views. Additionally, contrastive learning is leveraged between every two views to augment the view representations. During the multi-structure learning process, diverse clustering guidance is applied to view and view-fused representations, ultimately encouraging the clustering head to achieve an improved clustering structure. The hybrid ensemble clustering module comprises two essential components: attention-based ensemble and similarity-based ensemble. Specifically, the attention-based ensemble is focused on learning the contribution weights of various clustering structures. On the other hand, similarity-based ensemble leverages cluster assignment similarity and cluster classification dis-

similarity to guide the refinement of the consensus structure.

### 3.2. Multi-structure Processor

To learn multiple structures of different type representations, we design a multi-structure processor. Initially, view and view-fused representations are extracted during the representation learning phase. Subsequently, various clustering heads are employed on the representations to derive the clustering structure.

**Representation learning** A shared pre-trained semantic encoder, denoted as $\mathcal{E}(\cdot)$, is introduced to encode each document view. For the $i$-th document in the $v$-th view, $x_i^v$, its view representations $f_i^v \in \mathbb{R}^d$ are obtained as follows,

$$\begin{aligned} f_i^v &= h^v(z_i) = \mathsf{tanh}(W_h^v z_i^v + b_h^v) \\ z_i^v &= \mathcal{E}(x_i^v) \end{aligned} \tag{1}$$

where $h^v(\cdot)$ represents a multi-layer perceptron (MLP) layer with a hyperbolic tangent (tanh) activation function. $W_h^v$ and $b_h^v$ denote the weight and bias parameters of $h^v(\cdot)$, respectively.

To attain the desired separability among multi-view document samples while maintaining consistency within views, we incorporate contrastive learning as an aid to representation learning. Different view representations from the same sample are employed as positive pairs, while view representations from distinct samples are utilized as negative pairs. Formally, given a document sample $x_i$, positive pairs $(f_i^v, f_i^{\neg v})$ are generated for each view $v$ of $f_i^v$, in which $\neg v$ is one of the other document view except $v$. We set the negative pairs by pairing each view of document sample with all views from other document samples. The contrastive loss $\mathcal{L}_{rl}$ can be computed by

$$\mathcal{L}_{rl} = \frac{1}{2N} \sum_{(a,b)\in\mathcal{P}} -\log \frac{\exp(\mathsf{sim}(a,b)/\tau)}{\sum_{c \text{ s.t. } (a,c)\in\mathcal{N}} \exp(\mathsf{sim}(a,c)/\tau)} \tag{2}$$

where $N$ represents the number of document samples, $\mathcal{P}$ denotes the set of positive pairs, and $\mathcal{N}$ represents the set of negative pairs. In this context, we employ cosine similarity $\mathsf{sim}(\cdot)$, with $\tau$ serving as the temperature parameter to govern the level of softness. By default, we set $\tau = 1$.

During the representation learning stage, the minimization of $\mathcal{L}_{rl}$ facilitates the convergence of different view representations from the same sample, thus ensuring consistency among views, while concurrently driving apart the view representations of distinct samples, thereby enhancing separability between samples. After performing view representation learning, we generate a view-fused representation for each document sample by summing the individual view representations $\sum f_i^v$, denoted as $f_i^0$.

**Multi-structure learning** Given a document sample $x_i$, the representation learning process extracts $V$ view representations $\{f_i^v\}_{v=1}^V$ from each view and acquires an additional view-fused representation $f_i^0$. The objective of multi-structure learning is to discern distinct clustering structures from the ($V$+1) representations. In particular, diverse clustering heads $\{\zeta^v(\cdot)\}_{v=0}^V$ are employed to derive distinct clustering structures $\{A^v\}_{v=0}^V$ from representations $\{\{f_i^v\}_{i=1}^N\}_{v=0}^V$. Furthermore, distinct guidance mechanisms are employed to adjust the centroids of the clustering heads $\{\{\mu_k^v\}_{k=1}^K\}_{v=0}^V$ corresponding to the different representations.

The calculation of multiple clustering structures is accomplished using the clustering idea of DEC (Xie et al., 2016), applied to both view representations $\{f_i^v\}_{v=1}^V$ and view-fused representation $f_i^0$. For the sake of simplicity, we omit the superscripts representing views in the subsequent descriptions. Each element $a_{ik}$ of the cluster structure matrix $A \in \mathbb{R}^{N\times K}$ can be estimated using the Students' t-distribution as follows:

$$a_{ik} = \zeta(f_i, \mu_k) = \frac{(1 + ||f_i - \mu_k||^2)^{-1}}{\sum_{k'} (1 + ||f_i - \mu_{k'}||^2)^{-1}} \tag{3}$$

where, $a_{ik}$ is the $i$-th row and $k$-th column element of $A$, signifying the allocation probability of the $i$-th sample in the view to the $k$-th cluster.

Given the inherent inconsistency between the clustering structures embedded in the view representation and the view-fused representation, we have devised distinct clustering guidance for each of them. The aim is to refine the centroids of the clustering heads, ultimately achieving clustering allocation improvement.

1) **View structure guidance**: To emphasize representations assigned with high confidence and to mitigate the distortion of the clustering structure space caused by large clusters, we introduce the target structure $P \in \mathbb{R}^{N\times K}$, which is regarded as the target structure of the view cluster structure. The elements in $P$ can be estimated using:

$$p_{ik} = \frac{(a_{ik})^2 / \sum_{i'} a_{i'k}}{\sum_{k'} ((a_{ik'})^2 / \sum_{i'} a_{i'k'})} \tag{4}$$

where $p_{ik}$ is the $i$-th row and $k$-th column element of $P$, $a_{ik}$ is the element in clustering structure $A$.

Upon acquiring the target structure $P$, we employ the KL divergence between the view structure $A$ and the target structure $P$ as the learning objective to guide the optimization of the clustering head centroid $\mu_k$:

$$\mathcal{L}_{vs} = \sum_{v=1}^{V} \mathsf{KL}(P^v || A^v) = \sum_{v=1}^{V} \sum_i \sum_k p_{ik}^v \log \frac{p_{ik}^v}{a_{ik}^v} \tag{5}$$

2) **View-fused structure guidance**: The introduction of the view-fused structure learning objective

is motivated by the consideration of both representation space and structure space, aiming for an improved alignment with multi-view document clustering tasks. Specifically, we introduce the deep divergence-based clustering Loss (Kampffmeyer et al., 2019) to enhance the discriminative capacity of the learned structure. This loss comprises three terms, as presented below:

$$\mathcal{L}_{vfs} = \frac{1}{K}\sum_{i=1}^{K-1}\sum_{j>i}^{K}\frac{\alpha_i^T Q \alpha_j}{\sqrt{\alpha_i^T Q \alpha_i \alpha_j^T Q \alpha_j}} + triu(A^0 A^{0T}) + \frac{1}{K}\sum_{i=1}^{K-1}\sum_{j>i}^{K}\frac{m_i^T Q m_j}{\sqrt{m_i^T Q m_i m_j^T Q m_j}} \tag{6}$$

where $\alpha_i$ is the $i$-th column of the view-fused structure matrix $A^0$, $Q$ denotes the kernel similarity matrix estimated by $Q_{ij} = \exp(-||f_i^0 - f_j^0||^2/(2\sigma)^2)$, with $\sigma$ serving as the Gaussian kernel bandwidth, set to a default value of $0.15$. $triu(\cdot)$ refers to the strictly upper triangular, and $m_{ij} = \exp(-||\alpha_i - e_j||^2/(2\sigma)^2)$, where $e_j$ corresponds to corner $j$ of the standard simplex in $\mathbb{R}^K$.

In addition, we devise a self-generated margin center loss $\mathcal{L}_{mc}$ to facilitate the acquisition of cluster-relevant structures. This, in turn, directly contributes to the enhancement of the discriminative properties of the view-fused structure. $\mathcal{L}_{mc}$ is formulated as:

$$\mathcal{L}_{mc} = \sum_i \max(0, \Delta + ||f_i^0 - \mu_{y_i}^0|| - ||f_i^0 - \mu_{\neg y_i}^0||) \tag{7}$$

where $\mu_{y_i}^0$ is the $y_i$-th (the cluster label of view-fused structure $\alpha_i^0 \in A^0$) cluster center of semantic embedding $\{f_i^0\}_{i=1}^N$, $\mu_{\neg y_i}$ is the $\neg y_i$-th (a randomly selected cluster label other than $y_i$) cluster center, $\Delta$ represents the margin that controls the distance between intra- and inter-class pairs.

**Optimization** In an effort to ensure preservation of both consistency and complementarity between different views, we concurrently optimize the view structures and the view-fused structure as part of our overall objective, outlined as follows:

$$\mathcal{L}_{MsP} = \mathcal{L}_{vs} + \mathcal{L}_{vfs} + \mathcal{L}_{mc} \tag{8}$$

### 3.3. Hybrid Ensemble Clustering Module

Following the multi-structure learning process, we acquire a set of view structures and a view-fused structure. Subsequently, it becomes imperative to assess the distinct contributions of these various structures and integrate them into a consensus structure for the final clustering partition. To this end, we have developed a hybrid ensemble clustering module comprising the hybrid ensemble and self-supervised consensus clustering guidance, which facilitates the derivation of the final consensus structure from multiple clustering structures.

**Hybrid ensemble** Hybrid ensemble contains an attention-based structure ensemble and a similarity-based structure ensemble.

1) **Attention-based structure ensemble**: An attention-based structure ensemble is introduced to evaluate the different contributions of various view structures and integrate different structures into a consensus structure representation by self-attention mechanism.

Given a set of view structures and a view-fused structure $\{A^v\}_{v=0}^V$, we treat them as a sequential view input for the attention layer. The ensemble structure representation $E$ is estimated as follows:

$$E = \text{Reshape}(Att(\text{Stack}(\{A^v\}_{v=0}^V))) \tag{9}$$

where the $Att(\cdot)$ refers to a self-attention layer, and $\text{Stack}(\cdot)$ denotes the stack operation, where the corresponding output shape is $(N, (V+1), K)$. $\text{Reshape}(\cdot)$ changes the shape of $Att(\cdot)$'s output to $(N, (V+1) \times K)$. Through the process of attention-based learning, the ensemble structure representation $E$ captures both intra- and inter-view interactions.

2) **Similarity-based structure ensemble**: In order to incorporate the relationships between every pair of document samples, we have designed a similarity-based structure ensemble that considers both clustering assignment similarity and clustering assignment dissimilarity. More specifically, the clustering assignment similarity ensemble can be divided into two components, one dependent on soft cluster assignment and the other on hard cluster partition. The soft cluster assignment similarity ensemble matrix $M_s$ can be calculated using the clustering structures $\{A^v\}_{v=0}^V$:

$$\begin{aligned} M_s &= \tilde{A}\tilde{A}^T/K \\ \tilde{A} &= [A^0, A^1, \cdots, A^V] \end{aligned} \tag{10}$$

where the $[,]$ is the concatenated operation, $K$ is the number of clusters. The hard cluster partition similarity ensemble matrix $M_h$ can be estimated by the similar way:

$$\begin{aligned} M_h &= \bar{A}\bar{A}^T/K \\ \bar{A} &= [oh(A^0), oh(A^1), \cdots, oh(A^V)] \end{aligned} \tag{11}$$

where the $oh(\cdot)$ represents the one-hot function, which maps soft cluster assignment to hard cluster partition.

The pair-wise dissimilarity (Hussain et al., 2014) is also used as clustering assignment dissimilarity ensemble to combine the various structure. The pair-wise dissimilarity is calculated as the number

of differences in the multiple hard clustering partitions with respect to each document. And the dissimilarity matrix is transformed into a similarity matrix $M_{pd}$ by cosine similarity. As opposed to the similarity matrix $M_s, M_h$, the $M_{pd}$ takes into account not only how any two documents are assigned by the different clustering processes, but also their relationship with other classified documents.
Finally, we can simply get the final similarity-based structure ensemble:

$$M = M_s + M_h + M_{pd} \tag{12}$$

Note that, the $M$ can be used to get clustering partition directly.

**Consensus clustering guidance** In order to promote consensus structure learning in the final clustering partition space, an ensemble clustering head $g(\cdot)$ is built for attention-based ensemble structure representation $E$ to get its ensemble structure $A_e$ as follows:

$$A_e = g(E) \tag{13}$$

We still retain the KL divergence loss between the ensemble structure and its corresponding target structure as the main clustering guidance within the ensemble clustering module. By employing similar calculations in Eq.(4), we can obtain $P_e$ for ensemble structure $A_e$. The loss $\mathcal{L}_{kl}$ takes the following form:

$$\mathcal{L}_{kl} = \mathsf{KL}(P_e||A_e) \tag{14}$$

Taking into account that the clustering result with $K$ clusters can have $K$ factorial equivalent structures, we propose to gauge the disagreements between partitions by means of the similarities between samples as an auxiliary loss in consensus clustering. This approach is justified as clustering results are considered similar when their corresponding similarity matrices are in close proximity. The specific objective is as follows:

$$\mathcal{L}_{au} = \sum_{v=0}^{V} ||M - A^v A^{vT}||_F \tag{15}$$

where $M$ is the similarity-based structure ensemble matrix learned by Eq.(12), and $A^v$ is the view clustering structure matrix learned by multi-structure processor.

**Optimization** In this module, we design a dual-objective self-supervised mechanism, which unifies the hybrid ensemble in a uniform module. Therefore, the multiple structures can not only improve the separability of multi-view document data by learning a consensus structure, but also effectively feedback the view and view-fused representations and structures. The total loss is as follows:

$$\mathcal{L}_{HEC} = \mathcal{L}_{kl} + \lambda \mathcal{L}_{au} \tag{16}$$

where $\lambda > 0$ is a hyper-parameter that controls the optimization of the hybrid ensemble clustering module.

# 4. Experiments

## 4.1. Multi-view Document Datasets

We employed $4$ multi-view document datasets for conducting extensive experiments. Statistics of the datasets are summarized in Table 1.
The *BBC* dataset is derived from the BBC corpus which is a news article corpus originated from the BBC News. The BBC dataset contains $2,225$ news documents which are randomly selected from $5$ topical areas of BBC corpus, in particular, "business", "entertainment", "politics", "sports", and "technology". Each news document in the *BBC* dataset is described with the headline view and the content view.
We constructed the *HUFF* dataset to investigate the experimental performances of our proposed model on datasets with large number of clusters. The *HUFF* dataset is investigated by collecting $22,756$ timeline news blogs from $10$ topical areas of the HUFFPOST website[1]. Each news blog is represented in the headline and the short description view. The *HUFF-mini* dataset is a subset of the *HUFF* dataset, which contains $3,456$ blogs from $3$ distinguish topical areas, in particular, "Latino Voices", "Environment", "Education".
The *TOUTIAO* dataset is a newly generated news article dataset collected from the TouTiao News[2]. It contains $790$ documents organized in $4$ classes, in particular "Culture", "Military", "Technology", and "Sports". Each document is described by the headline view, the content view and converted author view.

Table 1: Summary of datasets.

| Dataset | No. of samples | No. of classes | Input length of each view |
|---|---|---|---|
| *BBC* | 2,225 | 5 | 24 & 288 |
| *HUFF-mini* | 3,456 | 3 | 120 & 120 |
| *HUFF* | 22,756 | 10 | 120 & 120 |
| *TOUTIAO* | 790 | 4 | 100 & 100 & 100 |

## 4.2. Experimental setting

we compared the proposed DMsECN with $16$ multi-view clustering models. Among them, we selected recent multi-view clustering models developed by traditional machine learning strategies that shows good clustering performances, including a series of models for multi-view subspace clustering (Brbić and Kopriva, 2018) (**P-MLRSSC**, **C-MLRSSC**, **P-KMLRSSC** and **C-KMLRSSC**), a

[1] https://www.huffpost.com
[2] http://www.toutiao.com

Table 2: Experimental results of multi-view document clustering on all experimental datasets (%). The best and second-best results are marked in **bold** and underline, respectively. Symbol "-" denotes no results are reported.

| Method | *BBC* | | | *HUFF-mini* | | | *HUFF* | | | *TOUTIAO* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI |
| P-MLRSSC | 87.26 | 66.80 | 71.30 | 62.33 | 35.59 | 28.78 | - | - | - | 71.75 | 60.46 | 53.00 |
| C-MLRSSC | 87.37 | 66.90 | 71.51 | 62.41 | 35.63 | 28.71 | - | - | - | 71.13 | 63.06 | 54.96 |
| P-KMLRSSC | 93.21 | 80.94 | 84.51 | 55.12 | 17.11 | 10.53 | - | - | - | 56.80 | 37.01 | 29.60 |
| C-KMLRSSC | 93.41 | 81.05 | 84.62 | 56.36 | 18.87 | 12.39 | - | - | - | 58.35 | 37.69 | 30.63 |
| MCDCF | 85.02 | 73.83 | 72.73 | 40.25 | 4.95 | 2.19 | - | - | - | 91.38 | 81.65 | 79.69 |
| FMR | 87.46 | 67.94 | 73.01 | 59.78 | 26.86 | 22.52 | - | - | - | 68.35 | 53.60 | 53.59 |
| MSC_IAS | 46.82 | 19.33 | 15.93 | 63.58 | 32.79 | 26.06 | 39.76 | 34.06 | 25.16 | 58.35 | 42.97 | 31.75 |
| SMVSC | 92.00 | 80.76 | 80.64 | 79.11 | 53.77 | 51.29 | 51.69 | 36.81 | 34.58 | 86.20 | 65.32 | 66.97 |
| APMC | 90.74 | 76.06 | 79.38 | 59.95 | 29.44 | 20.20 | 42.14 | 26.01 | 22.57 | 93.16 | 83.37 | 83.14 |
| SGF | 92.40 | 78.80 | 82.80 | 64.32 | 30.60 | 24.52 | 43.01 | 24.02 | 20.09 | 91.80 | 78.82 | 80.00 |
| DGF | 90.70 | 74.42 | 78.76 | 64.11 | 32.33 | 26.81 | 19.04 | 8.06 | 3.10 | 93.24 | 81.46 | 83.10 |
| MvDSCN | 45.53 | 23.26 | 20.07 | 52.78 | 16.49 | 13.91 | 40.01 | 26.70 | 23.11 | 79.62 | 53.71 | 54.07 |
| DEMVC | 62.11 | 35.90 | 35.68 | 77.08 | 40.33 | 46.74 | 42.95 | 26.85 | 24.09 | 85.06 | 68.42 | 68.90 |
| SURE | 75.60 | 48.42 | 50.27 | 54.98 | 12.54 | 12.15 | - | - | - | - | - | - |
| DealMVC | 64.63 | 46.93 | 44.48 | 61.69 | 34.41 | 36.38 | 52.72 | 36.91 | 33.95 | 76.84 | 59.03 | 51.97 |
| ProImp | 80.27 | 54.66 | 56.92 | 73.76 | 32.24 | 38.29 | 53.43 | 41.24 | 36.95 | 85.19 | 72.07 | 65.97 |
| DMsECN | **95.78** | **88.01** | **89.82** | **93.95** | **76.35** | **83.14** | **74.95** | **58.87** | **62.44** | **95.85** | **91.09** | **92.40** |

multi-view clustering model with deep concept factorization (Chang et al., 2021) (**MCDCF**), and some other multi-view clustering methods with promising performances (**FMR** (Li et al., 2019a), **MSC_IAS** (Wang et al., 2019), and **SMVSC** (Sun et al., 2021)). To investigating the performances of those models that make use of complementary and consistent information between the views, we also chose $3$ methods for comparison, in particular, **APMC** (Guo and Ye, 2019), **SGF** and **DGF** (Liang et al., 2019). Deep clustering models are also investigated for comparison, including autoencoder based models **MvDSCN** (Zhu et al., 2019) and **DEMVC** (Xu et al., 2021a) and contrastive learning based models **SURE** (Yang et al., 2022), **DealMVC** (Yang et al., 2023) and **ProImp** (Li et al., 2023).

We built our model on top of the pre-trained BERT model (with $12$-layer transformer) implemented in PyTorch (Wolf et al., 2019) and adopt most of its hyper-parameter settings. The input length for each view is shown in Table 1. If the view input exceeds the specified length, the view input will be truncated. To speed up the training process and avoid over-fitting, we forzen all the parameters of BERT except the last two transformer layers. For the dimension of the latent view representations, we set $d = 2K$ in our experiment. In the multi-structure processor, the representation learning and multi-structure learning both are conducted for 6-12 epochs with a $2e^{-5}$ learning rate by the Adam optimizer. Within the hybrid ensemble clustering module, training is conducted with $2e^{-}5 - 5e^{-5}$ learning rate by the Adam optimizer. To measure the performance of multi-view clustering methods, we employ $3$ metrics(Gan et al., 2007): Accuracy (ACC), Normalized Mutual Information(NMI), and Adjusted Rand Index (ARI) for evaluation. For all metrics, a higher value indicates better performance.

### 4.3. Comparison with State of the Arts

Table 2 presents the results achieved by our proposed model alongside those of other models. DMsECN outperforms all other models on all datasets by a substantial margin. The most significant performance enhancement was observed in the HUFF-mini dataset, with improvements of 22.58%, and 31.85% in NMI, and ARI, respectively. Even in the BBC dataset, which exhibited the least improvement, NMI, and ARI experienced improvements of 6.96%, and 5.20%, respectively. NMI places emphasis on information sharing, while ARI prioritizes data consistency. The improvements in NMI and ARI on each dataset exceed those in ACC, indicating that our proposed model excels in capturing the consistency and complementarity between different views, thereby enhancing both the separability and the compactness between clusters.

### 4.4. Ablation Study

**Multi-structure processor** Since the multi-structure processor is composed of representation learning and multi-structure learning, we conducted ablation studies on these two phases

Table 3: Clustering performance with multiple structures learning (%).

| $\mathcal{L}_{vs}$ | $\mathcal{L}_{vfs}$+$\mathcal{L}_{mc}$ | *BBC* | | | *HUFF-mini* | | | *HUFF* | | | *TOUTIAO* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI |
| | | 89.71 | 72.55 | 76.37 | 87.79 | 60.89 | 67.18 | 68.87 | 53.31 | 52.51 | 93.82 | 84.09 | 87.25 |
| ✔ | | 90.29 | 73.94 | 77.71 | 88.86 | 63.62 | 69.84 | 52.66 | 40.87 | 30.56 | 94.33 | 86.19 | 88.57 |
| | ✔ | 93.71 | 82.43 | 85.22 | 92.97 | 73.33 | 80.62 | 76.98 | 60.75 | 65.87 | 95.47 | 89.63 | 91.41 |
| ✔ | ✔ | 94.79 | 84.86 | 87.59 | 93.32 | 74.62 | 81.43 | 74.54 | 58.34 | 61.57 | 95.59 | 90.25 | 91.74 |

separately.

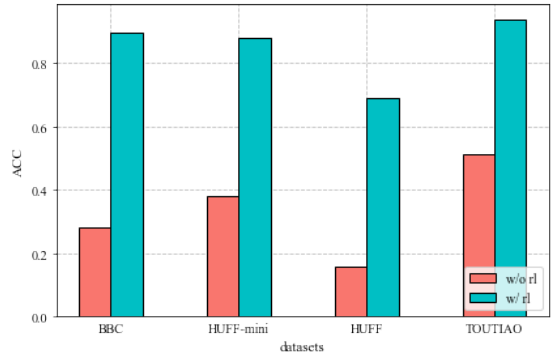



Figure 3: The multi-structure processor w/ or w/o representation learning.

1) Representation learning: The Figure 3 shows the comparison between with representation learning (w/ rl) and without representation learning (w/o rl) in multi-structure processor. From the figure, we can observe that although BERT is a pre-trained language model, it is still poor for semantic encoding of short text documents. However, by designing the contrastive loss for representation learning that fits with the multi-view document data, the semantics of multi-view document can already be mined. From the perspective of multi-structure learning, this representation learning based on contrastive loss can expand the variability between different samples, which is also useful for learning different clustering structures.
2) Multiple structure learning: To illustrate the necessity of multi-structure learning, we ablate view structure guidance and view fusion structure guidance in this part. Table 3 shows the experimental details of ablation. $\mathcal{L}_{vs}$ and $\mathcal{L}_{vfs}$+$\mathcal{L}_{mc}$ are clustering guidance corresponding to view structure learning and view fusion structure learning, respectively. Obviously, using $\mathcal{L}_{vs}$ can only get a little improvement (about 1%-3%). This may be because the clustering structure of the views is not uniform. In addition, since the view-fused guidance we designed introduces both the measurement of structure distribution, cluster-intra- and cluster-inter-distance, $\mathcal{L}_{vfs}$+$\mathcal{L}_{mc}$ can basically achieve similar performance to multi-structure learning. This phenomenon is particularly prominent on the *HUFF* dataset. In general, multi-structure learning comprehensively considers the optimization of views and view fusion structures, which can not only preserve the diversity between views, but also explore the underlying consistency in different views, and finally obtain a series of more robust clustering structures.

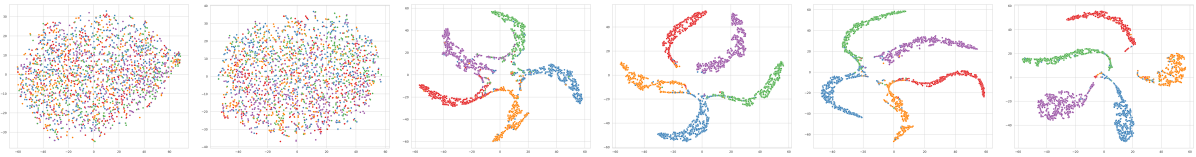
Table 4: The effectiveness of the hybrid ensemble clustering module. (%)

| Dataset | Metric | DMsECN w/o HECM | DMsECN |
|---|---|---|---|
| *BBC* | ACC | 94.79 | 95.78 |
| | NMI | 84.86 | 88.01 |
| | ARI | 87.59 | 89.82 |
| *HUFF-mini* | ACC | 93.32 | 93.95 |
| | NMI | 74.62 | 76.35 |
| | ARI | 81.43 | 83.14 |
| *HUFF* | ACC | 74.54 | 74.95 |
| | NMI | 58.34 | 58.87 |
| | ARI | 61.57 | 62.44 |
| *TOUTIAO* | ACC | 95.59 | 95.85 |
| | NMI | 90.25 | 91.09 |
| | ARI | 91.74 | 92.40 |

**Hybrid ensemble clustering module** We also conducted experiments for evaluating the effectiveness of the hybrid ensemble clustering module of our proposed DMsECN. Experimental results are depicted in Table 4. The "DMsECN w/o HECM" represents the proposed DMsECN without hybrid ensemble clustering module. It is clear that the introduction of hybrid ensemble clustering module achieves better performances than "DMsECN w/o HECM" on all datasets. Therefore, this module is useful for improving multi-view document clustering performance through hybrid ensemble. Specifically, refining ensemble structure with consensus clustering guidance is helpful for improving various clustering structures which promotes the clustering performance in return.

## 4.5. Visualization

In the Figure 4, subfigures (a) and (b) are the representation of headline view and content view ob-



(a) (b) (c) (d) (e) (f)

Figure 4: Visualization of different structure on the *BBC* dataset using t-SNE.

tained by original semantic learner. The subfigures (c) to (e) are the multiple clustering structure learned by multi-structure processor, respectively. In which, subfigure (e) is the view-fused structure. The subfigure (f) is the final consensus structure learned by hybrid ensemble clustering module. Compared with the structure of the original representation, the multi-structure processor has learned the basic cluster structure. Furthermore, the consensus structure obtained by hybrid ensemble clustering has clear cluster boundaries.

## 5. Conclusion

In this paper, we propose a deep multi-structure ensemble clustering network to cluster multi-view documents. The DMsECN consists of two essential components. The multi-structure processor extracts representations and clustering structures from multiple views. The hybrid ensemble clustering module is accountable for combining these clustering structures and generating a consensus structure that underpins the final clustering result. Our major contributions lie in the preservation of both consistency and complementarity between different views, and in the generation of a consensus that upholds cluster separability and compactness.

In future work, we will consider introducing graph convolutional networks (GCNs) in the stage of ensemble clustering to better capture the inter-sample structural relationships by mining neighborhood cluster structure of each view, thereby aiding multi-view document clustering tasks more effectively.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 62066007 and No. 62066008, the Key Technology R&D Program of Guizhou Province No. [2023]300 and No. [2022]277.

## References

Mahdi Abavisani and Vishal M Patel. 2018. Deep multimodal subspace clustering networks. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1601–1614.

Ruina Bai, Ruizhang Huang, Yanping Chen, and Yongbin Qin. 2021. Deep multi-view document clustering with enhanced semantic embedding. *Information Sciences*, 564:273–287.

Ruina Bai, Ruizhang Huang, Yongbin Qin, and Yanping Chen. 2022. Multi-view document clustering with joint contrastive learning. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, pages 706–719. Springer.

Maria Brbić and Ivica Kopriva. 2018. Multi-view low-rank sparse subspace clustering. *Pattern Recognition*, 73:247–258.

Shuai Chang, Jie Hu, Tianrui Li, Hao Wang, and Bo Peng. 2021. Multi-view clustering via deep concept factorization. *Knowledge-Based Systems*, 217:106807.

Ana LN Fred and Anil K Jain. 2005. Combining multiple clusterings using evidence accumulation. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):835–850.

Guojun Gan, Chaoqun Ma, and Jianhong Wu. 2007. Data clustering: Theory, algorithms, and applications.

Jun Guo and Jiahui Ye. 2019. Anchors bring ease: An embarrassingly simple approach to partial multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 118–125.

Hossein Hassani, Christina Beneki, Stephan Unger, Maedeh Taj Mazinani, and Mohammad Reza Yeganegi. 2020. Text mining in big data analytics. *Big Data and Cognitive Computing*, 4(1):1.

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

Dong Huang, Jianhuang Lai, and Chang-Dong Wang. 2016. Ensemble clustering using factor graph. *Pattern Recognition*, 50:131–142.

Dong Huang, Chang-Dong Wang, and Jian-Huang Lai. 2023. Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity. *IEEE Transactions on Knowledge & Data Engineering*, (01):1–16.

Dong Huang, Chang-Dong Wang, Hongxing Peng, Jianhuang Lai, and Chee-Keong Kwoh. 2018. Enhanced ensemble clustering via fast propagation of cluster-wise similarities. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1):508–520.

Syed Fawad Hussain, Muhammad Mushtaq, and Zahid Halim. 2014. Multi-view document clustering via ensemble method. *Journal of Intelligent Information Systems*, 43(1):81–99.

Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price. 2011. A link-based approach to the cluster ensemble problem. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2396–2409.

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

Guanzhou Ke, Zhiyong Hong, Wenhua Yu, Xin Zhang, and Zeyi Liu. 2022. Efficient multi-view clustering networks. *Applied Intelligence*, 52(13):14918–14934.

Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. 2023. Incomplete multi-view clustering via prototype-based imputation. In *Proceedings of the 32th International Joint Conference on Artificial Intelligence*.

Ruihuang Li, Changqing Zhang, Qinghua Hu, Pengfei Zhu, and Zheng Wang. 2019a. Flexible multi-view representation learning for subspace clustering. In *IJCAI*, pages 2916–2922.

Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, and Zhaohua Yang. 2019b. Deep adversarial multi-view clustering network. In *IJCAI*, pages 2952–2958.

Youwei Liang, Dong Huang, and Chang-Dong Wang. 2019. Consistency meets inconsistency: A unified graph learning framework for multi-view clustering. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1204–1209. IEEE.

Bingqian Lin, Yuan Xie, Yanyun Qu, Cuihua Li, and Xiaodan Liang. 2018. Jointly deep multi-view learning for clustering analysis. *arXiv preprint arXiv:1808.06220*.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. 2021. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11174–11183.

Xueying Niu, Chaowei Zhang, Xiaojie Zhao, Lihua Hu, and Jifu Zhang. 2023. A multi-view ensemble clustering approach using joint affinity matrix. *Expert Systems with Applications*, 216:119484.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617.

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

Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. 2017. From ensemble clustering to multi-view clustering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2843–2849.

Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. 2019. Marginalized multiview ensemble clustering. *IEEE transactions on neural networks and learning systems*, 31(2):600–611.

Alexander Topchy, Anil K Jain, and William Punch. 2005. Clustering ensembles: Models of consensus and weak partitions. *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1866–1881.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR.

Xiaobo Wang, Zhen Lei, Xiaojie Guo, Changqing Zhang, Hailin Shi, and Stan Z Li. 2019. Multi-view subspace clustering with intactness-aware similarity. *Pattern Recognition*, 88:50–63.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 478–487.

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

Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. 2021b. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9234–9243.

Xiaoqiang Yan, Yangdong Ye, Xueying Qiu, and Hui Yu. 2020. Synergetic information bottleneck for joint multi-view and ensemble clustering. *Information Fusion*, 56:15–27.

Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. 2022. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1055–1069.

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

Ming Yin, Weitian Huang, and Junbin Gao. 2020. Shared generative latent representation learning for multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6688–6695.

Changqing Zhang, Yeqing Liu, and Huazhu Fu. 2019. Ae2-nets: Autoencoder in autoencoder networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2577–2585.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

Xiaojia Zhao, Tingting Xu, Qiangqiang Shen, Youfa Liu, Yongyong Chen, and Jingyong Su. 2023. Double high-order correlation preserved robust multi-view ensemble clustering. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(1):1–21.

Runwu Zhou and Yi-Dong Shen. 2020. End-to-end adversarial-attention network for multi-modal clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14619–14628.

Pengfei Zhu, Binyuan Hui, Changqing Zhang, Dawei Du, Longyin Wen, and Qinghua Hu. 2019. Multi-view deep subspace clustering networks. *arXiv preprint arXiv:1908.01978*.