MASTER: A Multi-granularity Invariant Structure Clustering Scheme for Multi-view Clustering

Suixue Wang[1] , **Shilin Zhang**[2] , **Qingchen Zhang**[3,*] , **Peng Li**[4,5,*] and **Weiliang Huo**[1]
[1] School of Information and Communication Engineering, Hainan University
[2] College of Intelligence and Computing, Tianjin University
[3] School of Computer Science and Technology, Hainan University
[4] School of Computer Science and Technology, Dalian University of Technology
[5] Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology), Ministry of Education
{wangsuixue, wlhuo, zhangqingchen}@hainanu.edu.cn, zhang_shilin_sd@163.com, lipeng2015@mail.dlut.edu.cn

Abstract

Deep multi-view clustering has attracted increasing attention in the pattern mining of data. However, most of them perform self-learning mechanisms in a single space, ignoring the fruitful structural information hidden in different-level feature spaces. Meanwhile, they conduct the reconstruction constraint to learn generalized representations of samples, failing to explore the discriminative ability of complementary and consistent information. To address the challenges, a **M**ulti-granularity inv**A**riant **S**tructure clus**TER**ing scheme (MASTER) is proposed to define a bottom-up process that extracts multi-level information in sample, neighborhood, and category granularities from low-level, high-level, and semantics feature space, respectively. Specifically, it leverages the self-learning reconstruction with information-theoretic overclustering to capture invariant sample structure in the low-level feature space. Then, it models data diffusion of the clustering process in the reliable neighborhood to capture invariant local structure in the high-level feature space. Meanwhile, it defines dual divergences induced by the space geometry to capture invariant global structure in the semantics space. Finally, extensive experiments on 8 real-world datasets show that MASTER achieves state-of-the-art performance compared to 11 baselines.

1 Introduction

Multi-view clustering (MVC), as a fundamental task in machine learning, is attracting more and more prominent attention across various domains, such as image recognition and text classification, in the last few years [Wang *et al.*, 2024a; Zhou *et al.*, 2024; Jin *et al.*, 2023; Liu *et al.*, 2024a; Yu *et al.*, 2023]. It can be interpreted as an evolution of vanilla clustering for samples collected with multiple views. That is, MVC is capable of distilling complementary structure information of samples from multiple views/modalities to learn partition patterns in an unsupervised manner, such that samples within the same group are more similar to each other than samples from different groups [Chen *et al.*, 2023b; Cui *et al.*, 2024; Zou *et al.*, 2024; Wang *et al.*, 2024b].

With an encouraging renaissance of deep learning, some edge-cutting deep multi-view clustering methods have made significant progress in mining intrinsic patterns of data [Ren *et al.*, 2024; Gao *et al.*, 2024]. The deep multi-view clustering methods can be roughly divided into two groups, i.e., non-contrastive learning methods and contrastive learning methods, in accordance with self-learning mechanisms. The former draw on some vanilla clustering strategies, such as k-means and subspace clustering, as self-learning mechanisms to train neural networks, and they capture clustering-friendly complementary information of data via minimizing structure divergences between similar samples and maximizing structure divergences between dissimilar samples to model multi-view data partition. For example, [Xu *et al.*, 2023] introduces a self-supervised multi-view embedded clustering scheme to constrain structure consistencies between view-consensus features and view-specific features for MVC. The latter construct inter-view and intra-view contrastive learning mechanisms between data in multiple views to train their own multi-view fusion network, and they learn sample-invariant complementary information of data structure via pulling positive pairs of samples close and pushing negative samples apart to support the partition of multi-view data. For instance, [Yan *et al.*, 2023] introduces a structure-guided contrastive learning clustering scheme to weight similarities between samples in extracting invariant complementary information.

Although previous methods improve clustering performance by a significant margin, there still exist two challenges in mining intrinsic patterns of multi-view data. First, most of them perform self-learning mechanisms in a single space to extract data structure from multiple views for pattern mining, ignoring the fruitful multi-granularity structure information hidden in different-level feature spaces. Second, they learn generalized representations of samples in each view via the

*Corresponding authors: Qingchen Zhang and Peng Li.

reconstruction constraint to preserve consistent and complementary information, which causes degradations of the discriminative ability of view-specific representations. When confronted with multi-view data of complicated distributions, they may not produce desired clustering patterns.

To address the challenges, a **M**ulti-granularity inv**A**riant **S**tructure clus**TER**ing scheme (MASTER) is proposed to define a bottom-up clustering process, which captures fruitful structure information hidden in different-level feature space in a cascaded manner for multi-view data partition. In detail, MASTER consists of a sample-oriented representation learning strategy (SRL), a reliable neighborhood-driven diffusion strategy (RND), and a semantics-invariant cluster strategy (SIC). SRL leverages self-learning reconstructions with information-theoretic overclustering in encoding-decoding paradigms. It maximizes cross-view mutual information to perform sample clustering that interprets each sample as a single cluster in the low-level feature space, capturing invariant structure in the sample granularity for the enhancement of the discriminative ability. RND conducts a cross-view neighborhood distillation mechanism in the high-level feature space via imitating data diffusion of the clustering process. It introduces the cluster assignment information to dynamically refine neighbor relationships, learning invariant local structure in the neighborhood granularity. SIC defines dual divergences between assignment predictions of multi-view samples with simplexes of the semantics space. It leverages the space geometry to capture invariant global structure in the category granularity, maximizing inter-cluster separation and intra-cluster compactness for intrinsic pattern mining. Finally, extensive experiments are carried out on 8 benchmark datasets in comparison with 11 methods, and the results show MASTER achieves state-of-the-art performance.

Thus, the key contributions of the article can be concluded in the three-fold aspects.

- A multi-granularity invariant structure clustering scheme is proposed to define a bottom-up clustering process, which constructs a consistent hierarchy structure from the perspective of sample, neighborhood, and category to extract fruitful complementary and consistent information for MVC.
- MASTER imitates data diffusion in the clustering process to implement a cross-view neighborhood distillation mechanism, which can dynamically capture reliable affinity relationships in the neighborhood granularity to facilitate the extraction of invariant local structure from complementary information.
- Plenty of experiments are conducted on 8 real-world datasets and the results show MASTER achieves the state-of-the-art performance on multi-view data clustering in comparison with 11 methods.

2 Related Work

2.1 Non-contrastive Learning Methods

Non-contrastive learning methods introduce various self-learning mechanisms along with novel multi-view neural architectures, which measure structure divergences between samples, to extract clustering-friendly complementary information for clustering. Those methods can be divided into two-stage and one-stage methods.

The former train multi-view architectures with the aid of innovative self-learning mechanisms to learn clustering-friendly fusion representations, which perform additional clustering algorithms to divide multi-view samples [Zhou *et al.*, 2024]. For instance, [Li *et al.*, 2023] introduces an adversarial encoding-decoding architecture via defining an inter-view consistent cycle and an intra-view consistent cycle of data, which endows fusion representations of multi-view samples with clustering structure. [Dong *et al.*, 2023] introduces a multi-view encoder-decoder architecture for bipartite graphs, which aligns complementary information in each view to produce informative fusion representations of multi-view samples. The latter design effective self-learning mechanisms to train multi-view fusion neural architectures to produce clustering results in an end-to-end manner, which can cooperate structure discovery of data with complementary information fusion between views to promote sample partition [Xu *et al.*, 2024]. For example, [Zhou and Shen, 2020] designs a multi-view adversarial-attention fusion architecture, which leverages the geometry induced by simplexes of semantic space to guide divisions of multi-view data. [Gao *et al.*, 2024] proposes an adaptive multi-view semantics-invariant fusion network via constraining semantics alignments between views with samples, which further designs clustering as a Markov decision process of data partitions in an end-to-end manner.

2.2 Contrastive Learning Methods

Contrastive learning methods define cross-view contrastive learning mechanisms along with novel multi-view fusion architectures, which maximize similarities between positive samples and dissimilarities between negative samples, to learn view-invariant structure from views. Those methods can be also divided into two-stage and one-stage methods.

The former usually construct cross-view contrastive learning to model fusion features with view-invariant information of samples, and perform additional clustering constraints for pattern mining [Lu *et al.*, 2024]. For example, [Xu *et al.*, 2022] proposes a multi-level feature learning architecture via stacking contrastive learning networks on the middle layer of view-specific encoding-decoding networks, where cross-view consistencies of high-level features and semantic labels are constrained in a parallel manner for MVC. [Luo *et al.*, 2024] introduces a view-agnostic multi-view fusion contrastive architecture via interpreting the feature-dimensional concatenation between normalized data in each view as data-level fusion of samples, which conducts the vanilla contrastive learning on samples with the help of the manual noising/deleting data of samples in some views. The latter leverage cross-view contrastive learning to capture the distributions of data partition, which benefits from explorations of view-invariant information of samples in the semantic space [Zhang *et al.*, 2024b]. For instance, [Chen *et al.*, 2023a] proposes a cluster-level multi-view contrastive learning clustering method, which leverages view-invariant information via pulling positive pairs of clustering centroids close and push-

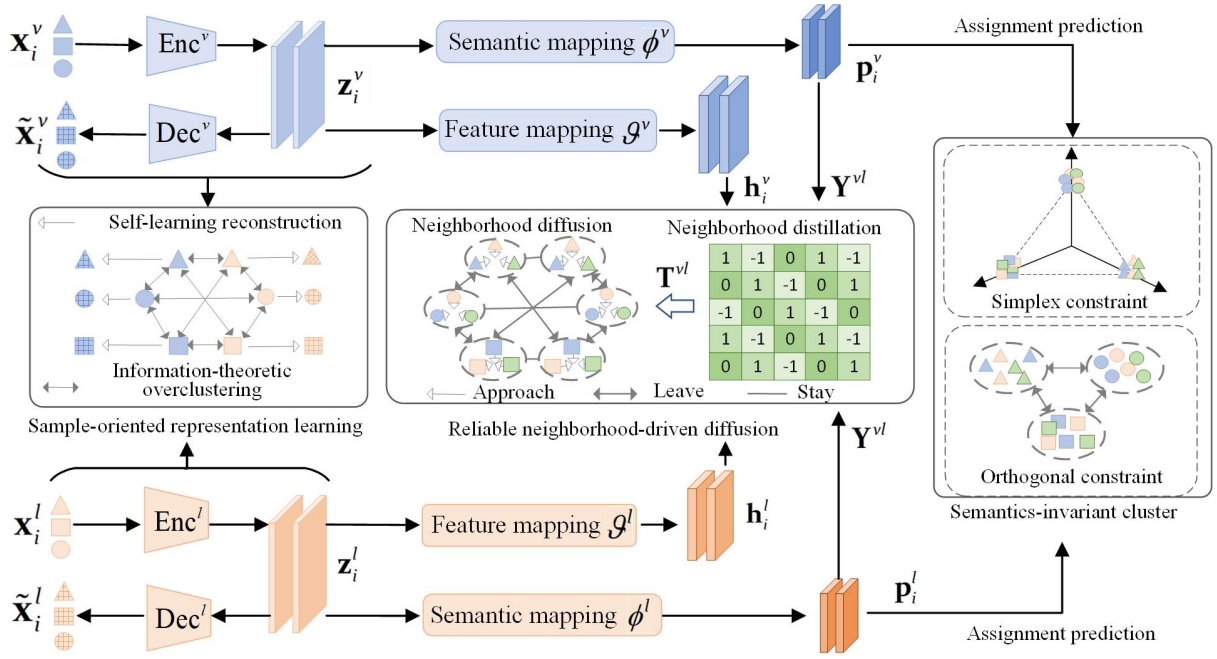


Figure 1: The flowchart of MASTER. **SRL**. Given a multi-view dataset $\{\mathbf{x}_i^v\}_{i=1,v=1}^{N,V}$, MASTER transforms data into the low-level feature space to obtain view-specific representations $\{\mathbf{z}_i^v\}_{i=1,v=1}^{N,V}$ within the encoder-decoder architecture, i.e., $\mathbf{z}_i^v = \mathrm{Enc}^v(\mathbf{x}_i^v)$ and $\tilde{\mathbf{x}}_i^v = \mathrm{Dec}^v(\mathbf{z}_i^v)$, and then leverages self-learning reconstructions with information-theoretic overclustering to capture invariant structure in the sample granularity. **RND**. MASTER learns the high-level representations $\{\mathbf{h}_i^v\}_{i=1,v=1}^{N,V}$ from $\{\mathbf{z}_i^v\}_{i=1,v=1}^{N,V}$ by the feature mapping function ϑ^v, i.e., $\mathbf{h}_i^v = \vartheta^v(\mathbf{z}_i^v)$. Subsequently, MASTER conducts a cross-view diffusion mask matrix $\mathbf{T}^{vl}$ via introducing the cross-view indicator matrix $\mathbf{Y}^{vl}$ into the neighborhood distillation mechanism, to learn invariant local structure in the neighborhood granularity. **SIC**. MASTER generates the assignment predictions $\{\mathbf{p}_i^v\}_{i=1,v=1}^{N,V}$ from $\{\mathbf{z}_i^v\}_{i=1,v=1}^{N,V}$ via the semantic mapping function ϕ^v, i.e., $\mathbf{p}_i^v = \phi^v(\mathbf{z}_i^v)$. Then, it leverages dual divergences between assignment predictions in the semantic space to capture invariant global structure in the category granularity.

ing negative pairs of clustering centroids apart. [Yang *et al.*, 2023] proposes a dual contrastive learning paradigm with an attention-weighted multi-view fusion encoder-decoder network, where pseudo-label graphs are used to calibrate global and local cross-view similarities between samples in end-to-end clustering, respectively.

3 Method

A multi-granularity invariant structure clustering scheme is proposed to define a bottom-up clustering process. It designs a sample-oriented representation learning strategy, a reliable neighborhood-driven diffusion strategy, and a semantics-invariant cluster strategy, to construct a consistent hierarchy structure from the perspective of sample, neighborhood, and category. The architecture of MASTER is shown in Figure 1.

3.1 The Sample-oriented Representation Learning Strategy

The sample-oriented representation learning strategy (SRL) leverages the invariant structure of each multi-view sample to enhance the discriminative ability of view-specific representations, which captures more complementary and consistent information from the low-level feature space in the sample granularity. To accomplish this, SRL constructs self-learning reconstructions within encoding-decoding paradigms in each view to model intrinsic view-specific representations from the data manifold. Then, SRL defines an information-theoretic overclustering that interprets each sample as an independent cluster, which endows view-specific representations with the discriminative invariant sample structure.

Specifically, given a multi-view dataset $\mathbf{X} = \{\mathbf{x}_i^v\}_{i=1,v=1}^{N,V}$ with C classes, which includes N samples with V views, SRL constructs an encoder-decoder pair that perform the self-learning reconstruction on data in each view, to capture view-specific representations $\mathbf{z}_i^v$, i.e., $\mathbf{z}_i^v = \mathrm{Enc}^v(\mathbf{x}_i^v)$ and $\tilde{\mathbf{x}}_i^v = \mathrm{Dec}^v(\mathbf{z}_i^v)$, which extracts generalized complementary and consistent information from data manifold. Thus, SRL constrains the self-learning reconstruction on data in the low-level feature space via

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{NV}\sum_{v=1}^{V}\sum_{i=1}^{N}||\mathbf{x}_i^v - \tilde{\mathbf{x}}_i^v||^2 \tag{1}$$

where $||\cdot||$ stands for the Euclidean norm and $\tilde{\mathbf{x}}_i^v$ is the i-th reconstruction sample in the v-th view. Eq. (1) can utilize the data manifold preservation to distill generalized complementary structure information for subsequent data partition.

Then, SRL maximizes the cross-view mutual information between view-specific representations to implement the information-theoretic overclustering, which endows view-specific representations with discriminative structure in the

sample granularity. In detail, given view-specific representations $\mathbf{z}_i^v$ and $\mathbf{z}_i^l$ in two views, SRL performs the information-theoretic overclustering via maximizing $\mathbb{I}(\mathbf{z}_i^v; \mathbf{z}_i^l)$, which is further transformed into the sample-level clustering on the basis of the data processing inequality, as follows:

$$\mathbb{I}(\mathbf{z}_i^v; \mathbf{z}_i^l) \geq \mathbb{I}(\mathbf{s}_i^v; \mathbf{s}_i^l), \tag{2}$$

where $\mathbb{I}(;)$ represents the mutual information. $\mathbf{s}_i^l$ denotes the probability vector of overclustering, and each element $\mathbf{s}_i^v[m] = \exp(\mathbf{z}_i^v[m])/\sum_{m'=1}^{M}\exp(\mathbf{z}_i^v[m'])(M > C)$. Then, SRL leverages the uniqueness concept of samples that each multi-view sample is a single cluster to maximize mutual information $\mathbb{I}(\mathbf{s}_i^v; \mathbf{s}_i^l)$.

$$\max \mathbb{I}(\mathbf{s}_i^v; \mathbf{s}_i^l) \Leftrightarrow \min \underbrace{-\mathbb{E}_{\mathbf{S}} \log \left[\frac{\frac{p(\mathbf{s}_i^v|\mathbf{s}_i^l)}{p(\mathbf{s}_i^v)}}{\frac{p(\mathbf{s}_i^v|\mathbf{s}_i^l)}{p(\mathbf{s}_i^v)} + \sum_{\mathbf{s}_j^v} \frac{p(\mathbf{s}_j^v|\mathbf{s}_i^l)}{p(\mathbf{s}_j^v)}} \right]}_{\mathcal{L}_{\mathrm{oc}}^{vl}(i)} \tag{3}$$

where $\mathbb{E}$ stands for the expectation operator. $\mathbf{S}$ denotes the set of $\mathbf{s}_i^v$. $p(\mathbf{s}_j^v|\mathbf{s}_i^l)/p(\mathbf{s}_j^v)$ measures the influence of knowing $\mathbf{s}_i^l$ on $\mathbf{s}_j^v$, $j \neq i$. SRL constrains the information-theoretic overclustering on the multi-view dataset as

$$\mathcal{L}_{\mathrm{oc}} = \sum_{i=1}^{N}\sum_{v=1}^{V}\sum_{l\neq v}^{V} \mathcal{L}_{\mathrm{oc}}^{vl}(i) \tag{4}$$

Thus, SRL leverages the structural information of each sample to enhance the discriminative ability of generalized representations via

$$\mathcal{L}_{\mathrm{SRL}} = \mathcal{L}_{\mathrm{rec}} + \alpha\mathcal{L}_{\mathrm{oc}} \tag{5}$$

where α is a trade-off parameter.

3.2 The Reliable Neighborhood-driven Diffusion Strategy

The reliable neighborhood-driven diffusion strategy (RND) focuses on extracting cross-view invariant structure from complementary information in the neighborhood granularity to promote local consistency of affinity relationship in the high-level feature space for multi-view sample partition. To achieve this, RND defines a cross-view neighborhood distillation mechanism that imitates inherent paradigms of data diffusion in the clustering process, i.e., approaching, leaving, and staying, to dynamically learn an invariant structure-enhanced neighbor relationship with the help of the cluster assignment information. In detail, RND models the approaching paradigm for the in-neighborhood multi-view samples with the same cluster assignment, to enhance the intra-neighborhood consistency of data structure. RND models the leaving paradigm for the out-of-neighborhood multi-view samples with different cluster assignments, to strengthen the inter-neighborhood divergences of data structure. Meanwhile, it models the staying paradigm for the multi-view samples that do not satisfy the above two conditions, to ensure the robustness of the data structure. RND learns reliable neighbor relationships via modeling three diffusion paradigms in the high-level feature space to enhance the exploration of invariant local structure.

Specifically, given the view-specific representations $\mathbf{Z}^v = \{\mathbf{z}_i^v\}_{i=1}^N$ of the v-th view in the low-level feature space, RND conducts a feature mapping function ϑ^v to generate the high-level representations $\mathbf{H}^v = \{\mathbf{h}_i^v\}_{i=1}^N$, i.e., $\mathbf{h}_i^v = \vartheta^v(\mathbf{z}_i^v)$. Subsequently, RND introduces the cross-view indicator matrix $\mathbf{Y}^{vl} \in \mathbb{R}^{N\times N}$ of cluster assignments to implement the neighborhood distillation mechanism. The mechanism adaptively refines neighbor relationships to generate a diffusion mask matrix $\mathbf{T}^{vl} \in \mathbb{R}^{N\times N}$ between the v-th view and the l-th view:

$$\begin{aligned} \mathbf{t}_{ij}^{vl} &= \begin{cases} 1 & \text{if } \mathbf{h}_j^l \in \psi^l(\mathbf{h}_i^v),\ \mathbf{y}_{ij}^{vl} = 1 \\ 0 & \text{if } \mathbf{h}_j^l \notin \psi^l(\mathbf{h}_i^v),\ \mathbf{y}_{ij}^{vl} = 0 \\ -1 & \text{otherwise} \end{cases} \\ \mathbf{y}_{ij}^{vl} &= \begin{cases} 1 & \text{if } \mathbf{y}_i^v = \mathbf{y}_j^l \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

where $\psi^l(\mathbf{h}_i^v)$ denotes the K-nearest neighbor set of $\mathbf{h}_i^v$ in the l-th view. $\mathbf{y}_{ij}^{vl} = 1$ indicates that two samples are assigned to the same cluster ($\mathbf{y}_i^v = \mathbf{y}_j^l$), vice versa. In Eq. (6), $\mathbf{t}_{ij}^{vl} = 1$ indicates that in-neighborhood samples with the same cluster assignment should approach each other. $\mathbf{t}_{ij}^{vl} = 0$ indicates that the out-of-neighborhood samples with different cluster assignments should leave each other. $\mathbf{t}_{ij}^{vl} = -1$ indicates that the sample does not satisfy the above two conditions should stay stationary.

Thus, RND utilizes the neighborhood distillation mechanism induced by the data diffusion, to dynamically maximize intra-neighborhood similarities and minimize inter-neighborhood similarities in the high-level feature space, which is formulated as follows:

$$\mathcal{L}_{\mathrm{RND}} = \sum_{v=1}^{V}\sum_{l\neq v}^{V} \mathcal{H}(\mathbf{T}^{vl}, \mathbf{H}^v(\mathbf{H}^l)^\top) + \mathcal{H}(\mathbf{T}^{vl}, \mathbf{H}^v(\mathbf{H}^v)^\top) \tag{7}$$

where $\mathcal{H}$ denotes the cross entropy function.

3.3 The Semantics-invariant Cluster Strategy

The semantics-invariant cluster strategy (SIC) aims to capture the invariant global structure of the semantics space from complementary and consistent information in the category granularity, which ensures inter-cluster separation and intra-cluster compactness for intrinsic pattern mining. To this end, SIC defines dual divergences, i.e., the simplex constraint and orthogonal constraints, between the assignment predictions of multi-view samples with the simplexes of the semantics space in two dimensions, which leverages the geometry of the semantics space to guide explorations of invariant global structure in the category granularity.

To be specific, given the view-specific representations $\mathbf{Z}^v = \{\mathbf{z}_i^v\}_{i=1}^N$ of multi-view data in the v-th view, SIC defines a semantic mapping function ϕ^v from view-specific representations of samples to the assignment predictions $\mathbf{P}^v = \{\mathbf{p}_i^v\}_{i=1}^N$, i.e., $\mathbf{p}_i^v = \phi^v(\mathbf{z}_i^v)$, which can leverage the discriminative complementary and consistent information of each view to capture invariant global structure of multi-view data.

Algorithm 1 MASTER

Input: The multi-view dataset $\mathbf{X} = \{\mathbf{x}_i^v\}_{i=1,v=1}^{N,V}$
Parameter: Learnable parameters in encoders Enc^v, decoders Dec^v, feature mapping functions ϑ^v, and semantic mapping functions ϕ^v of multiple views
Output: Clustering partition $\overline{\mathbf{P}}$

while not converged **do**
  Sample structure learning:
  Extract view-specific representations by encoder-decoder pairs $\mathbf{z}_i^v = \mathrm{Enc}^v(\mathbf{x}_i^v)$ and $\tilde{\mathbf{x}}_i^v = \mathrm{Dec}^v(\mathbf{z}_i^v)$.
  Compute probability vectors of overclustering by $\mathbf{s}_i^v[m] = \exp(\mathbf{z}_i^v[m]) / \sum_{m'=1}^{M} \exp(\mathbf{z}_i^v[m'])$.
  Neighborhood structure learning:
  Generate high-level representations $\mathbf{h}_i^v = \vartheta^v(\mathbf{z}_i^v)$.
  Generate diffusion mask matrices $\mathbf{T}^{vl}$ by Eq. (6).
  Global structure learning:
  Obtain assignment predictions $\mathbf{p}_i^v = \phi^v(\mathbf{z}_i^v)$.
  Update learnable parameters by minimizing Eq. (5), Eq. (7), and Eq. (10).
end while
return Clustering partition $\overline{\mathbf{P}} = \frac{1}{V}\sum_{v=1}^{V} \mathbf{P}^v$

Afterwards, SIC interprets the C simplexes in the C-dimensional semantics space as the cluster centroids, and it minimizes the divergence between samples and the corresponding cluster centroid via the cross entropy, which can push multi-view samples within the same cluster towards the same simplex to capture clustering-friendly invariant structure of multi-view data. SIC computes the loss induced by the centroids with C dimensions as follows:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\sum_{v=1}^{V} \mathbf{c}_i \log \mathbf{p}_i^v \tag{8}$$

$$\mathbf{c}_{ik} = \begin{cases} 1 & \text{if } \max_r \overline{\mathbf{p}}_{ir} = \overline{\mathbf{p}}_{ik} \\ 0 & \text{otherwise} \end{cases} \quad r = 1, 2 \cdots C, \overline{\mathbf{p}} = \frac{1}{V}\sum_{v=1}^{V} \mathbf{p}^v$$

where $\mathbf{c}_{ik}$ is the k-th element of the i-th simplex of the semantics space, and $\overline{\mathbf{p}}_{ik}$ denotes the k-th element in the average assignment prediction vector of the i-th multi-view sample. In such a manner, SIC pushes samples within the same cluster towards the same simplex and samples of the different clusters apart, maximizing the inter-cluster separation and intra-cluster compactness.

Then, SIC recasts each column of the assignment prediction matrix in each view as the representations of cluster centroids in the sample dimension, and it forces inter-view and intra-view orthogonal constraints between the centroids, which can further maximize cross-view consistencies to capture the invariant global structure across views. SIC computes the loss induced by the centroids with N dimensions with the simplex via contrastive learning as follows:

$$\mathcal{L} = \sum_{l=1}^{V}\sum_{v=1}^{V} \left\| (\mathbf{P}^l)^{\mathrm{T}} \mathbf{P}^v - \mathbf{I} \right\|^2 \tag{9}$$

where $\mathbf{I}$ denotes the identity matrix with $C \times C$ dimensions.

Dataset	Class	Sample	View	Type
BDGP	5	2500	2	Vector
CCV	20	6773	3	Vector
MNIST-USPS	10	5000	2	Image
Fashion-MV	10	10000	3	Image
Caltech-2V	7	1440	2	Vector
Caltech-3V	7	1440	3	Vector
Caltech-4V	7	1440	4	Vector
Caltech-5V	7	1440	5	Vector

Table 1: The detailed statistics of the 8 multi-view datasets

Eq. (9) constrains inter-view consistencies and intra-view compactness of multi-view sample assignments. Thus, SIC captures the invariant structure of the semantics space in the category granularity via:

$$\mathcal{L}_{\mathrm{SIC}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{v=1}^{V} \mathbf{c}_i \log \mathbf{p}_i^v + \sum_{l=1}^{V}\sum_{v=1}^{V} \left\| (\mathbf{P}^l)^{\mathrm{T}} \mathbf{P}^v - \mathbf{I} \right\|^2 \tag{10}$$

The overall loss of MASTER is shown in Eq. (11). Algorithm 1 lists the detailed steps of MASTER.

$$\mathcal{L} = \mathcal{L}_{\mathrm{SRL}} + \beta \mathcal{L}_{\mathrm{RND}} + \gamma \mathcal{L}_{\mathrm{SIC}} \tag{11}$$

where β and γ are trade-off parameters.

4 Experiment

4.1 Experiment Settings

Benchmark Datasets: Plenty of experiments are conducted on 8 real-world datasets to validate the performance of MASTER. In detail, MNIST-USPS is a handwritten digit dataset that contains 5000 bi-view samples within 10 classes. BDGP is a Drosophila embryo dataset that contains 2500 bi-view samples within 5 classes. CCV is a video dataset with 6773 tri-view samples belonging to 20 classes. Fashion is an image dataset about products, which contains 10000 tri-view samples within 10 classes. Caltech-2V, Caltech-3V, Caltech-4V, and Caltech-5V are created from the Caltech dataset that consists of RGB images with multiple views. The statistics of the 8 datasets are listed in Table 1.

Comparison Methods: 11 methods with state-of-the-art performance are used to verify the superiority of MASTER, which are composed of non-contrastive learning methods and contrastive learning methods. The former consist of SDMVC [Xu *et al.*, 2023], MVCAN [Xu *et al.*, 2024], DMAC-SI [Gao *et al.*, 2024], CAMVC [Zhang *et al.*, 2024a], DSMVAGC [Wang *et al.*, 2024a], and AEVC [Liu *et al.*, 2024b]. The latter include DealMVC [Yang *et al.*, 2023], GCFAgg [Yan *et al.*, 2023], SCM [Luo *et al.*, 2024], CSOT [Zhang *et al.*, 2024b], and DIVIDE [Lu *et al.*, 2024]. In the experiments, the comparison methods are implemented according to settings in the original papers, and the grid search is utilized on the trade-off parameters suggested by the authors to guarantee the best performance of the comparison methods.

Implementation Details: MASTER consists of V encoder-decoder pairs, a feature mapping network, and a seman-

Method	BDGP			CCV			MNIST-USPS			Fashion-MV		
	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
DealMVC(2023)	0.9600	0.9197	0.9069	0.2935	0.2882	0.1426	0.9778	0.9615	0.9765	0.9636	0.9522	0.9636
SDMVC(2023)	0.9816	0.9447	0.9548	0.3125	0.3085	0.3364	0.9880	0.9815	0.9880	0.8626	0.9215	0.8405
GCFAgg(2023)	0.9870	0.9624	0.9870	0.3543	0.3292	0.3812	0.9956	0.9871	0.9956	0.9758	0.9674	0.9758
MVCAN(2024)	0.9732	0.9287	0.9732	0.3269	0.3111	0.3622	0.9818	0.9778	0.9818	0.9205	0.9489	0.9217
DMAC-SI(2024)	0.9880	0.9634	0.9723	0.3012	0.2998	0.3185	0.9882	0.9826	0.9882	0.8857	0.9247	0.8623
CAMVC(2024)	0.9540	0.8746	0.9540	0.2709	0.2656	0.2987	0.8098	0.7022	0.8098	0.7897	0.7817	0.7886
SCM(2024)	0.9710	0.9130	0.9557	0.2562	0.2353	0.2909	0.9882	0.9672	0.9747	0.9347	0.9072	0.9347
DSMVAGC(2024)	0.8932	0.8372	0.8932	0.2337	0.1908	0.2679	0.8289	0.7543	0.8593	0.7902	0.7605	0.7938
AEVC(2024)	0.7126	0.4513	0.7125	0.1640	0.1132	0.1874	0.8181	0.6967	0.8189	0.8257	0.7779	0.8235
DIVIDE(2024)	0.9514	0.9039	0.9514	0.3101	0.3012	0.3373	0.9676	0.9244	0.9676	0.9290	0.8791	0.9292
CSOT(2024)	0.9896	0.9655	0.9896	0.3167	0.3098	0.3468	0.9924	0.9787	0.9924	0.9754	0.9663	0.9754
MASTER	**0.9920**	**0.9727**	**0.9920**	**0.3805**	**0.3441**	**0.4159**	**0.9964**	**0.9892**	**0.9964**	**0.9916**	**0.9790**	**0.9916**
	Caltech-2V			Caltech-3V			Caltech-4V			Caltech-5V		
	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
DealMVC(2023)	0.5207	0.4289	0.3054	0.5871	0.5606	0.4398	0.7580	0.6952	0.6470	0.8407	0.7651	0.7208
SDMVC(2023)	0.5350	0.4508	0.5893	0.6786	0.6049	0.6943	0.7543	0.6888	0.7741	0.8789	0.7825	0.8789
GCFAgg(2023)	0.6643	0.5008	0.6643	0.6400	0.5345	0.6529	0.7343	0.6610	0.7343	0.8336	0.7331	0.8336
MVCAN(2024)	0.5698	0.5017	0.5988	0.6807	0.6037	0.6971	0.7392	0.6609	0.7862	0.8914	0.8077	0.8907
DMAC-SI(2024)	0.6087	0.5323	0.6333	0.7008	0.6312	0.7098	0.7998	0.7276	0.7998	0.8802	0.7850	0.8802
CAMVC(2024)	0.6345	0.4864	0.6345	0.7114	0.5959	0.7150	0.8414	0.7272	0.8414	0.8821	0.7988	0.8821
SCM(2024)	0.6271	0.4708	0.6271	0.6500	0.5186	0.6836	0.7143	0.5788	0.7143	0.8055	0.7298	0.8041
DSMVAGC(2024)	0.5371	0.4537	0.5312	0.5898	0.5547	0.5977	0.6039	0.5717	0.6022	0.6644	0.5724	0.6426
AEVC(2024)	0.5743	0.4633	0.5843	0.6795	0.5673	0.6886	0.7604	0.5664	0.7604	0.7564	0.6731	0.7564
DIVIDE(2024)	0.5822	0.5294	0.5822	0.6089	0.5375	0.6077	0.6432	0.5793	0.6552	0.7515	0.6829	0.7515
CSOT(2024)	0.6390	0.5440	0.6420	0.7050	0.6230	0.7150	0.8120	0.6900	0.8120	0.7910	0.7126	0.7910
MASTER	**0.6900**	**0.5875**	**0.6900**	**0.7307**	**0.6446**	**0.7450**	**0.8432**	**0.7600**	**0.8440**	**0.9124**	**0.8255**	**0.9124**

Table 2: The numerical results on the 8 datasets in comparison with 11 state-of-the-art methods in terms of ACC, NMI, and PUR.

tic mapping network, which are implemented via fully-connected layers with the following architectures: D-500-500-2000-256, 256-2000-500-500-D, 256-128, and 256-C, respectively. D and C represent the dimensionality of the original data and semantic features, respectively. MASTER optimizes the overall networks via the Adam optimizer with a learning rate of 0.0003 across all datasets. In the optimization, the number of training epochs is set to 350. The batch size is set to 256. The values of α, β, and K are determined via the ablation experiments with the grid search strategy on each dataset. γ is set to 1 on all the datasets.

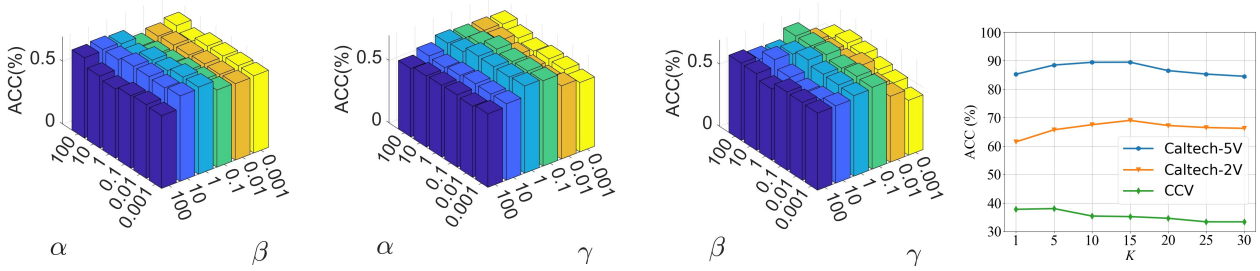
Evaluation Metrics: Three standard clustering metrics, i.e., accuracy (ACC), normalized mutual information (NMI), and purity (PUR) are utilized to verify the superior performance of MASTER, and the larger value of the metrics indicates the better clustering performance. In the experiments, the results shown in Tables **??**, Table 3, and Figure 2, are the average of 5 runs for all the metrics to guarantee a fair comparison.

4.2 Clustering Performance

Table **??** demonstrates the numerical results produced on the 8 datasets. Those results show that MASTER outperforms all the comparison methods by a significant margin in terms of the three evaluation metrics, which can be concluded in three observations below. ① BDGP, MNIST-USPS, and Fashion-MV: The clustering results are very high and close to the maximum value 1, meaning that there is not much room for improvement on the three datasets. That is, the multi-view samples, which are assigned into incorrect clusters by the second-best method, are very hard to distinguish in pattern mining. In such a condition, MASTER can still assign more samples into the correct clusters and achieve better performance in terms of all the evaluation metrics on the three datasets. ② CCV and Caltech-2V: The clustering results are relatively low, meaning that it is hard to recognize some multi-view samples. That is, the multi-view samples that are often assigned to incorrect clusters by the second best method, are very hard to distinguish in pattern mining. In such a condition, MASTER can still assign more samples into the correct clusters and achieve better performance in terms of all the evaluation metrics on the three datasets. ③ Caltech-3V, Caltech-4V, and Caltech-5V: The clustering results are of significant gains, indicating that MASTER is competent to capture intrinsic patterns from the complementary and consistent information of multi-view data. In detail, MASTER outperforms all the comparison methods more and more significantly as the number of views increases. Especially, MASTER outperforms the second-best method by 2.10%, 1.78%, and 2.17% in terms of ACC, NMI, and PUA on Caltech-5V.

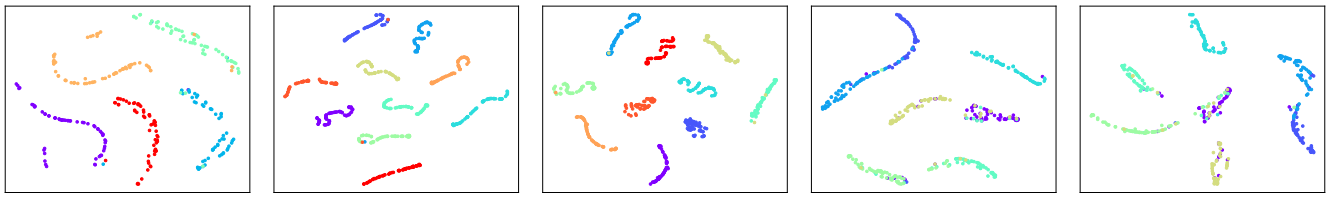
4.3 The Ablation and Hyper-parameter Analyses

Ablation Results: Table 3 demonstrates the ablation results on MNIST-USPS and Caltech-2V, which is utilized to evaluate the effectiveness and rationality of MASTER in terms of ACC and NMI. As shown in Table 3, there are three variants. MASTER w/o SRL, MASTER w/o RND, and MASTER w/o



(a) (b) (c) (d)

Figure 2: The results of the hyper-parameter analysis. (a), (b), and (c) show the results of α, β, γ on Caltech-2V. (d) demonstrates the results on the in-neighborhood sample number K in RND.



(a) BDGP (b) MNIST-USPS (c) Fashion-MV (d) Caltech-4V (e) Caltech-5V

Figure 3: The clustering result visualizations of MASTER on 5 datasets.

	MNIST-USPS		Caltech-2V	
	ACC	NMI	ACC	NMI
MASTER w/o SRL	0.9852	0.9618	0.6644	0.5500
MASTER w/o RND	0.9854	0.9842	0.6323	0.5273
MASTER w/o SIC	0.9928	0.9795	0.6799	0.5819
MASTER	**0.9964**	**0.9892**	**0.6900**	**0.5875**

Table 3: The ablation results on MNIST-USPS and Caltech-2V in terms of ACC and NMI.

SIC represent removal of SRL, RND, and SIC, respectively. Two observations can be concluded from Table 3. ① The three variants yield inferior results on both datasets. That is, the removal of each component causes a decrease in clustering performance, which validates the effectiveness of each component. ② MASTER outputs the best clustering results on both datasets in terms of ACC and NMI, which further validates the design rationality of MASTER.

Hyper-parameter Results: Figure 2 presents the influence of the trade-off parameters α, β, and γ on Caltech-2V in terms of ACC. In detail, all trade-off parameters are set in $\{100, 10, 1, 0.1, 0.01, 0.001\}$. Then, one parameter is fixed and the other two are adjusted in the hyper-parameter analyses. As shown in the results of Figure 2, MASTER is robust to the selection of trade-off parameters. That is, when α, β, and γ are in $\{10, 1, 0.1\}$, $\{10, 1, 0.1, 0.01\}$, and $\{1, 0.1, 0.01\}$, respectively, MASTER can produce stable results close to the optimal result within a large range of settings. In addition, Figure 2(d) illustrates the influence of the selection of in-neighborhood sample number K in RND, and the results show that MASTER is of high robustness and can find the optimal number of in-neighborhood samples.

4.4 The Visualization Results

Figure 3 visualizes the clustering results that are carried out on 5 datasets to further evaluate the clustering superiority of MASTER. In detail, 1000 multi-view samples are randomly selected from each of 5 datasets, which are further input into MASTER to produce assignment predictions. Then, the t-SNE algorithm is conducted on the assignment predictions to output the visualization of clustering results.

As shown in Figure 3, multi-view samples belonging to the same cluster are densely clustered within the same region, and there are no overlaps between distinct clusters. This observation substantiates that MASTER is competent to mine patterns that ensure the inter-cluster separation and intra-cluster compactness.

5 Conclusion

In this paper, a multi-granularity invariant structure clustering scheme is proposed to define a bottom-up process, which extracts multi-level information in sample, neighborhood, and category granularities. Specifically, MASTER captures invariant sample structure in the low-level feature space, invariant local structure in the high-level feature space, and invariant global structure in the semantics space, in which the three components cooperate to enhance the discriminative ability of the complementary and consistent information for MVC. Finally, extensive experiments on 8 real-world datasets verify that MASTER achieves state-of-the-art performance in comparison with 11 baselines.

Acknowledgments

This study was supported by Hainan Provincial Natural Science Foundation of China (Grant No. 825CXTD608) and grants (No. 62162023 and No. KYQD(ZR)-21079).

References

[Chen *et al.*, 2023a] Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16706–16715, Paris, France, October 2023. IEEE.

[Chen *et al.*, 2023b] Zhe Chen, Xiaojun Wu, Tianyang Xu, and Josef Kittler. Fast self-guided multi-view subspace clustering. *IEEE Trans. Image Process.*, 32:6514–6525, 2023.

[Cui *et al.*, 2024] Jinrong Cui, Yuting Li, Han Huang, and Jie Wen. Dual contrast-driven deep multi-view clustering. *IEEE Trans. Image Process.*, 33:4753–4764, 2024.

[Dong *et al.*, 2023] 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 19383–19394, Paris, France, October 2023. IEEE.

[Gao *et al.*, 2024] Jing Gao, Meng Liu, Peng Li, Jianing Zhang, and Zhikui Chen. Deep multiview adaptive clustering with semantic invariance. *IEEE Trans. Neural Networks Learn. Syst.*, 35(9):12965–12978, September 2024.

[Jin *et al.*, 2023] Shan Jin, Zhikui Chen, Shuo Yu, Muhammad Altaf, and Zhenchao Ma. Self-augmentation graph contrastive learning for multi-view attribute graph clustering. In *Proceedings of the 2023 Workshop on Advanced Multimedia Computing for Smart Manufacturing and Engineering*, pages 51–56, 2023.

[Li *et al.*, 2023] Peng Li, Asif Ali Laghari, Mamoon Rashid, Jing Gao, Thippa Reddy Gadekallu, Abdul Rehman Javed, and Shoulin Yin. A deep multimodal adversarial cycle-consistent network for smart enterprise system. *IEEE Trans. Ind. Informatics*, 19(1):693–702, 2023.

[Liu *et al.*, 2024a] Han Liu, Junjie Sun, Xiaotong Zhang, and Hongyang Chen. New intent discovery with multi-view clustering. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12381–12385. IEEE, 2024.

[Liu *et al.*, 2024b] Suyuan Liu, Ke Liang, Zhibin Dong, Siwei Wang, Xihong Yang, Sihang Zhou, En Zhu, and Xinwang Liu. Learn from view correlation: An anchor enhancement strategy for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26151–26161, Seattle, USA, June 2024. IEEE.

[Lu *et al.*, 2024] 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 38th AAAI Conference on Artificial Intelligence*, pages 14193–14201, Vancouver, Canada, February 2024. AAAI Press.

[Luo *et al.*, 2024] Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. Simple contrastive multi-view clustering with data-level fusion. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 4697–4705, Jeju, South Korea, Augus 2024. ijcai.org.

[Ren *et al.*, 2024] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S. Yu, and Lifang He. Deep clustering: A comprehensive survey. *Trans. Neural Networks Learn. Syst.*, Early Access:1–21, 2024.

[Wang *et al.*, 2024a] Siwei Wang, Xinwang Liu, Suyuan Liu, Wenxuan Tu, and En Zhu. Scalable and structural multi-view graph clustering with adaptive anchor fusion. *IEEE Trans. Image Process.*, 33:4627–4639, 2024.

[Wang *et al.*, 2024b] Suixue Wang, Huiyuan Lai, Shuling Wang, and Qingchen Zhang. Contramae: Contrastive alignment masked autoencoder framework for cancer survival prediction. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2621–2626. IEEE, 2024.

[Xu *et al.*, 2022] 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 16030–16039, New Orleans, USA, June 2022. IEEE.

[Xu *et al.*, 2023] Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, Philip S. Yu, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Trans. Knowl. Data Eng.*, 35(7):7470–7482, 2023.

[Xu *et al.*, 2024] Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, and Xiaofeng Zhu. Investigating and mitigating the side effects of noisy views for self-supervised clustering algorithms in ppractical multi-view scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22957–22966, Seattle, USA, June 2024. IEEE.

[Yan *et al.*, 2023] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19863–19872, Vancouver, Canada, June 2023. IEEE.

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

[Yu *et al.*, 2023] Sanshi Lei Yu, Qi Liu, Fei Wang, Yang Yu, and Enhong Chen. Federated news recommendation with

fine-grained interpolation and dynamic clustering. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3073–3082, 2023.

[Zhang *et al.*, 2024a] Chao Zhang, Xiuyi Jia, Zechao Li, Chunlin Chen, and Huaxiong Li. Learning cluster-wise anchors for multi-view clustering. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 16696–16704, Vancouver, Canada, February 2024. AAAI Press.

[Zhang *et al.*, 2024b] Qian Zhang, Lin Zhang, Ran Song, Runmin Cong, Yonghuai Liu, and Wei Zhang. Learning common semantics via optimal transport for contrastive multi-view clustering. *IEEE Trans. Image Process.*, 33:4501–4515, 2024.

[Zhou and Shen, 2020] Runwu Zhou and Yi-Dong Shen. End-to-end adversarial-attention network for multi-modal clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14607–14616, Seattle, USA, June 2020. IEEE.

[Zhou *et al.*, 2024] Lihua Zhou, Guowang Du, Kevin Lü, Lizhen Wang, and Jingwei Du. A survey and an empirical evaluation of multi-view clustering approaches. *ACM Comput. Surv.*, 56(7):187:1–187:38, 2024.

[Zou *et al.*, 2024] Guoliang Zou, Yangdong Ye, Tongji Chen, and Shizhe Hu. Learning dual enhanced representation for contrastive multi-view clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8731–8739, Melbourne, Australia, October 2024. ACM.