INCOMPLETE MULTI-VIEW CLUSTERING VIA INFERENCE AND EVALUATION

Binqiang Huang, Zhijie Huang, Shoujie Lan, Qinghai Zheng, Yuanlong Yu**

College of Computer and Data Science, Fuzhou University, China

ABSTRACT

Multi-view clustering aims to improve the clustering performance by leveraging information from multiple views. Most existing works assume that all views are complete. However, samples in real-world scenarios cannot be always observed in all views, leading to the challenging problem of Incomplete Multi-View Clustering (IMVC). Although some attempts are made recently, they still suffer from the following two limitations: (1) they usually adopt shallow models, which are unable to sufficiently explore the consistency and complementary of multiple views; (2) they lack of a suitable measurement to evaluate the quality of the recovered data during the learning process. To address the aforementioned limitations, we introduce a novel Incomplete Multi-View Clustering via Inference and Evaluation (IMVC-IE). Specifically, IMVC-IE adopts the contrastive learning strategy on features of different views to excavate the underlying information from existing samples firstly. Subsequently, massive alternative simulated data are inferred for missing views and a novel evaluation strategy is presented to obtain the proper data for missing views completion. Extensive experiments are conducted and verify the effectiveness of our method.

***Index Terms*—** Incomplete multi-view clustering, missing data inference, data evaluation

1. INTRODUCTION

Existing Incomplete Multi-View Clustering (IMVC) methods can be roughly categorized into two types, *i.e.*, traditional and deep methods [1]. Traditional IMVC methods leverage utilize zero or mean values to complete the missing views [2], and then use, for example, non-negative matrix decomposition based methods [3], subspace learning based methods [4], kernel learning based methods [5], and graphical methods [6] [7] to perform multi-view clustering. However, traditional IMVC methods have limited representation capabilities and are difficult to handle high complexity problems [8]. In recent years, deep IMVC methods have gradually attracted attention due to their strong generalization ability and scalability [9]. Deep IMVC methods usually design some kind of filling strategy to infer missing data before clustering, and then get results based on the recovered data [10, 11, 12].

Despite the remarkable progress, most existing traditional IMVC methods suffer from limitations of the shallow model and naive data completion, namely, zero padding and mean padding. For Deep IMVC methods, they ignore the evaluation of the recovered data, which may impede the improvement of clustering performance. To address these limitations, a novel method, termed Incomplete Multi-View Clustering via Inference and Evaluation (IMVC-IE), is introdued in this paper. To be specific, IMVC-IE has two components, *i.e.*, Data Inference Module (DI) and Data Evaluation Module (DE).

Unlike the direct zero-padding or mean-padding, in which the populated data are often far from the real data, DI presents a novel strategy based on the assumption that: when samples are sufficient, the data distribution converges to the overall distribution. Therefore, it is possible to infer with high confidence that the missing data fluctuates around the data mean. In other words, if we generate a large number of simulated truth data near the view data mean, the inference of missing data can be accomplished by combining appropriate evaluation methods to determine which simulated truth data is more likely to be the true data. Obviously, DI is a statistical learning approach for missing data inference.

Regarding DE, the relationship between views is considered here. Specifically, inspired by the work of CLIP [13], we first design a deep auto-encoder to filter the intra-view noise [14] and map the non-missing paired data from each view to a common similarity space, then the contrastive loss of different views is employed to effectively excavated the underlying relationships, which is utilized to for data evaluation. Based on DI and DE, the proposed method can complete the missing views effectively. Based on the recovered multi-view data, we get the clustering results by exploring the consistency in semantic space. Main contributions are as follows:

- By applying the Data Inference (DI) and the Data Evaluation (DE) modules, the proposed method combines statistical learning and deep learning. Therefore, the missing views can be recovered with high confidence in our method.
- Based on the recovered multi-view data, the clustering results are achieved by pursuing the consistency in the semantic space via conservative learning. Experimental results verify the effectiveness and competitiveness of our method.

*Corresponding authors: Qinghai Zheng and Yuanlong Yu. This work was partially supported by the National Natural Science Foundation of China (62306074, U21A20471), the Natural Science Foundation of Fujian Province (2023J05025), the Youth Innovation Funds of Fujian Province (JAT220005), and the Research Startup Project of Fuzhou University (XRC-22033).

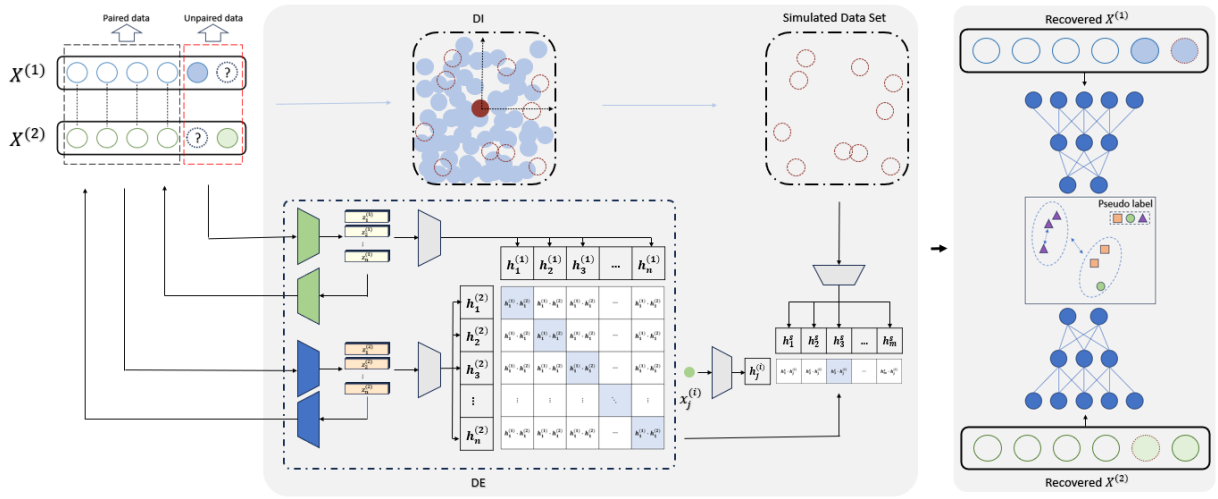

Fig. 1. General flowchart of our algorithm. Incomplete data are first fed into Data Inference Module (DI) and Data Evaluation Module (DE) in parallel, where the former is responsible for inferencing simulated data sets based on the mean and standard deviation of the dataset, and the latter trains a data evaluator using the existing complete paired data. Afterwards, the generated simulated data sets and the corresponding data from other views are fed into the data evaluator to identify the best simulated data. After recovering the missing data one by one, the complete data is feed into the semantic clustering module.

2. METHOD

Notation: Considering multi-view data $\{X = X^{(v)}\}_{v=1}^{V}$, $X^v \in \mathbb{R}^{d_v \times N}$, where V is the number of views, N denotes the number of samples, and d_v indicates the feature dimension of the v-th view, we divide the data into two parts: paired data that exists in all views, and unpaired data that exists only in some partial views.

2.1. Data Inference Module

When the dataset has sufficient samples of the overall population, the data distribution converges to the overall distribution, hence the missing data fluctuates around $\bar{x}$, $\bar{x}$ is the data mean. Specifically, we measure the "range of fluctuation" as the standard deviation of each feature dimension of the data, i.e., $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, ..., \sigma_{d_v}\}$.In the process of generating the *j*-th simulated true data of *i*-th data of *v*-th view, the generated sample is denoted as $\widetilde{x}_{ij}^{(v)}$.

Based on the above analysis, the Data Inference (DI) module is introduced to generate massive alternative samples for missing views. To be specific, we first calculate the *v*-th view data mean $\bar{x}^{(v)}$ and the set of standard deviations for each feature dimension $\boldsymbol{\sigma}^{(v)} = \{\sigma_1^{(v)}, \sigma_2^{(v)}, ..., \sigma_{d_v}^{(v)}\}$, and then generate a Gaussian noise with zero mean and standard deviation equal to $\boldsymbol{\sigma}^{(v)}$ which is denoted as e and can be calculated. Finally, $\widetilde{x}_{ij}^{(v)} = \bar{x}^{(v)} + e$ form the simulated truth data. In order to allow the simulated truth data to radiate to the real data at a high confidence level, a collection of simulated truth data with a size of not less than half of the dataset is generated.

2.2. Deep Evaluation Module

The evaluator is designed to select the simulated truth data with the highest similarity to the complementary view paired data. In order to fully explore the underlying information and learn the compact suitable view-specific feature representations. we adopt a two-level multi-view feature representation [14], which is specified by first extracting the low-level features from the original dataset using an automatic coder, and then using contrast learning to catch high-level features focusing on the public semantics of the views on top of the low-level features.

Specifically, for the *m*-th view, we denote the encoder and decoder by $E^{(m)}(X^m; \theta_E^{(m)})$ and $D^{(m)}(X^m; \theta_D^{(m)})$, where $\theta_E^{(m)}$ and $\theta_D^{(m)}$ denote the network parameters, denote $z_i^{(m)} = E^{(m)}(x_i^{(m)}) \in R^{d_z}$ as the d_z-dimensional latent features of the *i*-th sample, denote $\hat{x}_i^{(m)} = D^{(m)}(z_i^{(m)})$ as the reconstruction of the *i*-th sample, and denote $\mathcal{L}_Z^{(m)}$ as the reconstruction loss of the input $X^{(m)}$ and output $\hat{X}^{(m)}$. Therefore, we can construct the loss term $\mathcal{L}_Z = \sum_{m=1}^{M} \mathcal{L}_Z^{(m)}$ as follows:

$$\sum_{m=1}^{M} \mathcal{L}_Z^{(m)} = \sum_{m=1}^{M} \sum_{i=1}^{N} \left\| x_i^{(m)} - D^{(m)}(E^{(m)}(x_i^{(m)})) \right\|_2^2. \quad (1)$$

Since the feature $\{Z^{(m)}\}_{m=1}^{M}$ mixes view public information with view private information, we treat $\{Z^{(m)}\}_{m=1}^{M}$ as a low-level feature and learn another level of features, i.e., high-level features $\{H^{(m)}\}_{m=1}^{M}$, We stack the feature MLP on $\{Z^{(m)}\}_{m=1}^{M}$ to obtain high-level features $\{H^{(m)}\}_{m=1}^{M}$, where

$h_i^{(m)} \in R^{d_H}$ and the feature MLP is a one-layer linear MLP denoted by $F(\{Z^{(m)}\}_{m=1}^M; W_H)$. In the low-level feature space, we utilize reconstruction goals to maintain the representational power of $\{Z^{(m)}\}_{m=1}^M$ and thus avoid the problem of model collapse. In the high-level feature space, we further achieve the consistency goal through contrast learning so that $\{H^{(m)}\}_{m=1}^M$ focuses on learning the common semantics of all views. Specifically, each high-level feature $h_i^{(m)}$ has $(MN-1)$ feature pairs, i.e., $\{h_i^{(m)}, h_j^{(n)}\}_{j=1,\ldots,N}^{n=1,\ldots M}$, where $\{h_i^{(m)}, h_i^{(n)}\}_{m \neq n}$ is an $(M-1)$ positive sample pair and the remaining $M(N-1)$ feature pairs are negative sample pairs. In contrast learning, the similarity of positive sample pairs should be maximized and the similarity of negative sample pairs should be minimized. Inspired by NT-Xent [15], cosine distance is applied to measure the similarity between two features. To be clear, it is formulated as follows:

$$d(h_i^{(m)}, h_j^{(n)}) = \frac{\langle h_i^{(m)}, h_j^{(n)} \rangle}{\left\| h_i^{(m)} \right\| \left\| h_j^{(n)} \right\|}, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is the dot product operator. Then, the feature contrastive loss between $H^{(m)}$ and $H^{(n)}$ is formulated as:

$$l_{fc}^{(mn)} = -\frac{1}{N} \sum_{i=1}^{N} log \frac{e^{d(h_i^{(m)}, h_j^{(n)})/\tau_F}}{\sum_{j=1}^{N} \sum_{v=m,n} e^{d(h_i^{(m)}, h_j^{(v)})/\tau_F} - e^{1/\tau_F}}. \tag{3}$$

where τ_F denotes the temperature parameter. In this paper, we design an accumulated multi-view feature contrastive loss across all views as follows:

$$\mathcal{L}_H = \frac{1}{2} \sum_{m=1}^{M} \sum_{n \neq m} l_{fc}^{(mn)}. \tag{4}$$

Up to this point, once the training is complete, for a given input, end-to-end similarity evaluation can be done with the paired data corresponding to another viewpoint. For example, the set of simulated data and the paired data corresponding to the other view are fed into the evaluator, and the simulated data with the highest cosine similarity to the paired data is selected as the best truth data.

2.3. Clustering based on the recovered data

Semantic consistency contrast learning is introduced in order to effectively mine the variable semantic consistency information in the semantic space. Based on the fact that multiple views describe the same goal, we introduce a shared classifier $C(\cdot)$ with parameter φ. The last layer of the classifier network is *softmax*. By using the classifier, we map $\{X^{(m)}\}_{m=1}^M$ to a semantic space of dimension size *k*, where *k* is the number of categories in the multi-view dataset. Due to the diversity of specific statistical information of multiple views, the semantic information of different views in the semantic space may be confusing, which leads to diverse and confusing results for $\{X^{(m)}\}_{m=1}^M$. Therefore, we constrain that $\{X^{(m)}\}_{m=1}^M$ should have similar pseudo-labels. We introduce contrast learning to mine consistent semantic

information in the semantic space while obtaining consistent categories. Thus

$$Q^{(m)} = C(E^{(m)}(X^{(m)}; \varphi)) \in R^{N \times k}. \tag{5}$$

As same as the *Evaluation* part, let $q_i^{(m)} \in R^k$, then the cosine distance between $q_i^{(m)}$ and $q_j^{(n)}$ is formulated as:

$$d(q_i^{(m)}, q_j^{(n)}) = \frac{\langle q_i^{(m)}, q_j^{(n)} \rangle}{\left\| q_i^{(m)} \right\| \left\| q_j^{(n)} \right\|}. \tag{6}$$

The semantic contrast loss is formulated as:

$$l_{sc}^{(mn)} = -\frac{1}{N} \sum_{i=1}^{N} log \frac{e^{d(q_i^{(m)}, q_j^{(n)})/\tau_Q}}{\sum_{j=1}^{N} \sum_{v=m,n} e^{d(q_i^{(m)}, q_j^{(v)})/\tau_Q} - e^{1/\tau_F}}, \tag{7}$$

where τ_Q is the temperature parameter. We design the cumulative multi-view semantic contrast loss on all views as:

$$\mathcal{L}_Q = \frac{1}{2} \sum_{m=1}^{M} \sum_{n \neq m} l_{sc}^{(mn)} + \sum_{m=1}^{M} \sum_{j=1}^{K} s_j^m log s_j^m, \tag{8}$$

where $s_j^m = \frac{1}{N} \sum_{i=1}^{N} q_{ij}^m$, adding this regularization term helps to avoid assigning all samples to a single cluster [16].

3. EXPERIMENTS

In order to validate the effectiveness of our method, a number of clustering experiments are conducted in this section. Important statistics are summarized in Table 1 and Fig. 3, a brief introduction is presented below.

3.1. Experimental setup

Datasets: **BDGP** is a benchmark dataset with two views. One is a visual view and another one is a textual view. It contains 2,500 images about Drosophila embryos belonging to 5 categories. Each image is represented by 1,750-D visual vectors and 79-D textual feature vectors. **MNIST-USPS** is a popular dataset of handwritten digits containing 5,000 samples featuring two different styles of digit images comprising two different viewpoints, with each handwritten digit represented by a 784-D visual vector. **UCI** dataset consists of features of handwritten numerals ("0" - "9") extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of 2,000 patterns) have been digitized in binary images.

Baselines: We compare the proposed IMVC-IE with several state-of-the-art multi-view clustering algorithms, namely **IMVTSC [17]**, **UEAF [18]**, **HCLS_CGL [19]**. Meanwhile, naive IMVC methods based on the **Zero-padding** and **Mean-padding** are also employed as comparison algorithms here.

Table 1. Clustering results in the metric of ACC of different methods with different missing rates (0.5, 0.7, and 0.9). The bold numbers indicate the best clustering performance.

datasets	BDGP			MNIST-USPS			UCI		
missing rate	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
IMVTSC	35.04	32.80	24.92	36.64	26.96	14.22	50.10	37.25	16.60
UEAF	52.72	45.04	27.60	49.52	29.98	14.02	51.15	38.10	16.65
HCLS_CGL	47.40	34.64	23.88	67.52	46.20	16.78	64.65	41.95	17.40
Mean-padding	35.52	35.06	21.66	33.02	28.59	13.24	41.87	43.25	15.40
Zero-padding	32.94	27.68	16.28	28.42	21.90	13.36	41.50	27.38	16.28
Ours	**64.90**	**55.18**	**58.78**	**71.70**	**60.65**	**50.44**	**64.88**	**58.58**	**51.68**

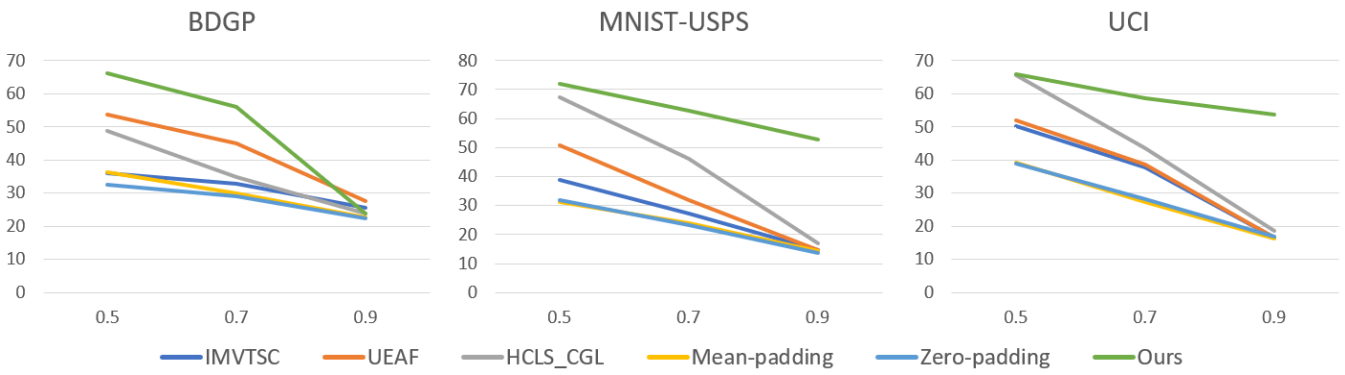

Fig. 2. Clustering results in the metric of purity of different methods with different missing rates (0.5, 0.7, and 0.9).

Table 2. Ablation experimental results in the metric of ACC at a missing rate of 0.5 for each data set.

datasets	BDGP	MNIST-USPS	UCI
missing rate	0.5	0.5	0.5
Without Inference	49.48	42.13	48.16
Without Evaluation	49.20	42.10	47.81
IMVC-IE	**69.90**	**71.70**	**64.88**

To quantitatively show the clustering performance of different methods, the clustering results are presented in metrics of ACCuarcy (ACC) and purity in this section.

3.2. Clustering results and ablation experiments

The incomplete multi-view clustering performance with various missing rates, *i.e.,* 0.5, 0.7, and 0.9, can be found in Table 1. In general, IMVC-IE achieves the promising clustering performance in all cases. As shown in Table 1, the best clustering results can be achieved by our method with different missing rates in the metric of ACC. For example, around 4.18%, 14.45%, and 33.66% improvements can be obtained in MNIST-USPS with missing rates fixed to 0.5, 0.7, and 0.9. Although UEAF gets the slightly better result than our method in BDGP with 0.9 missing rate in the metric of purity, the performance of our method is also competitive. Furthermore, our IMVC-IE achieves the considerable improvements in other cases. To verify the effectiveness, we set up two kinds of ablation experiments, the first one removes the inference module, i.e., random noise is used to fill the missing data. The second one removes the evaluation module, i.e., Interpolation is randomly selected from the set of simulated truth data. We conducted the experiments with a 0.5 missing rate on each dataset. The experimental results in the metric of ACC are shown in Table 2, which verifies the effectiveness of DI and DE introduced in our method.

4. CONCLUSION

We propose a novel incomplete multi-view clustering method, termed IMVC-IE. To overcome the limitations of most existing IMVC methods, two modules, *i.e.,* Data Inference (DI) and Data Evaluation (DE) are introduced in our IMVC-IE. By generating massive alternative samples for missing views based on DI from the perspective of statistic learning, a proper sample can be selected by DE to complete the missing view by investigating the underlying relationships between views. Experimental results on three real-world benchmark datasets demonstrate the effectiveness of our algorithm.

5. REFERENCES

[1] Jie Xu, Chao Li, Yazhou Ren, Liang Peng, Yujie Mo, Xiaoshuang Shi, and Xiaofeng Zhu, "Deep incomplete multi-view clustering via mining cluster complementarity," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 8761–8769.

[2] Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang, and Jinxing Li, "A survey on incomplete multiview clustering," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 2, pp. 1136–1149, 2022.

[3] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou, "Partial multi-view clustering," in *Proceedings of the AAAI conference on artificial intelligence*, 2014, vol. 28.

[4] Shiping Wang, Zhaoliang Chen, Shide Du, and Zhouchen Lin, "Learning deep sparse regularizers with applications to multi-view clustering and semi-supervised classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5042–5055, 2021.

[5] Xinwang Liu, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu, "Efficient and effective regularized incomplete multi-view clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 8, pp. 2634–2646, 2020.

[6] Jie Wen, Ke Yan, Zheng Zhang, Yong Xu, Junqian Wang, Lunke Fei, and Bob Zhang, "Adaptive graph completion based incomplete multi-view clustering," *IEEE Transactions on Multimedia*, vol. 23, pp. 2493–2504, 2020.

[7] Zhaoliang Chen, Lele Fu, Jie Yao, Wenzhong Guo, Claudia Plant, and Shiping Wang, "Learnable graph convolutional network and feature fusion for multi-view learning," *Information Fusion*, vol. 95, pp. 109–119, 2023.

[8] Jun Guo and Jiahui Ye, "Anchors bring ease: An embarrassingly simple approach to partial multi-view clustering," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 118–125.

[9] Jie Wen, Chengliang Liu, Shijie Deng, Yicheng Liu, Lunke Fei, Ke Yan, and Yong Xu, "Deep double incomplete multi-view multi-label learning with incomplete labels and missing views," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[10] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu, "Generative partial multi-view clustering with adaptive fusion and cycle consistency," *IEEE Transactions on Image Processing*, vol. 30, pp. 1771–1783, 2021.

[11] Cai Xu, Ziyu Guan, Wei Zhao, Hongchang Wu, Yunfei Niu, and Beilei Ling, "Adversarial incomplete multi-view clustering.," in *IJCAI*, 2019, vol. 7, pp. 3933–3939.

[12] Jie Wen, Zhihao Wu, Zheng Zhang, Lunke Fei, Bob Zhang, and Yong Xu, "Structural deep incomplete multi-view clustering network," in *Proceedings of the 30th ACM international conference on information & knowledge management*, 2021, pp. 3538–3542.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[14] 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*, 2022, pp. 16051–16060.

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

[16] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool, "Scan: Learning to classify images without labels," in *European conference on computer vision*. Springer, 2020, pp. 268–285.

[17] Jie Wen, Zheng Zhang, Zhao Zhang, Lei Zhu, Lunke Fei, Bob Zhang, and Yong Xu, "Unified tensor framework for incomplete multi-view clustering and missing-view inferring," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 10273–10281.

[18] Jie Wen, Zheng Zhang, Yong Xu, Bob Zhang, Lunke Fei, and Hong Liu, "Unified embedding alignment with missing views inferring for incomplete multi-view clustering," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 5393–5400.

[19] Jie Wen, Chengliang Liu, Gehui Xu, Zhihao Wu, Chao Huang, Lunke Fei, and Yong Xu, "Highly confident local structure based consensus graph learning for incomplete multi-view clustering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15712–15721.