Imputation-free Incomplete Multi-view Clustering via Knowledge Distillation

Benyu Wu, **Wei Du***, **Jun Wang** and **Guoxian Yu***
School of Software, Shandong University, Jinan, China
bywu109@163.com, {duwei, kingjun, gxyu}@sdu.edu.cn

Abstract

Incomplete multi-view data presents a significant challenge for multi-view clustering (MVC). Existing incomplete MVC solutions commonly rely on data imputation to convert incomplete data into complete data. However, this paradigm suffers from the risk of error accumulation when clustering unreliably imputed data, causing suboptimal clustering performance. Moreover, using imputation to fulfill missing data is inefficient, while inferring data categories based solely on the existing views is extremely challenging. To this end, we propose an **I**mputation-free **I**ncomplete **MVC** (`I2MVC`) via pseudo-supervised knowledge distillation. Specifically, `I2MVC` decomposes the incomplete MVC problem into two tasks: an MVC task for complete data and a pseudo-supervised classification task for fully incomplete data. A self-supervised simple contrastive Teacher network is trained for clustering complete data, and its knowledge is distilled into a lightweight pseudo-supervised Student network. The Student network, unrestricted by view completeness, further guides the clustering of fully incomplete data. Finally, the clustering results from both tasks are merged to generate the final clustering outcome. Experimental results on benchmark datasets demonstrate the effectiveness of `I2MVC`.

1 Introduction

Multi-view clustering (MVC) [Bickel and Scheffer, 2004; Yao *et al.*, 2019; Wei *et al.*, 2021; Yu *et al.*, 2024] is an unsupervised learning paradigm specifically devised for the analysis of multi-view data (MVD), which clusters data by exploiting the feature similarities across different views. Multi-view data, which describes objects from multiple perspectives, is ubiquitous in real-world scenarios. For example, in medical diagnosis, diseases can be examined using computed tomography, magnetic resonance imaging, and ultrasound imaging. By harnessing the consistency and complementarity of information across multiple views, more precise diagnostic results can be obtained.

*Corresponding authors

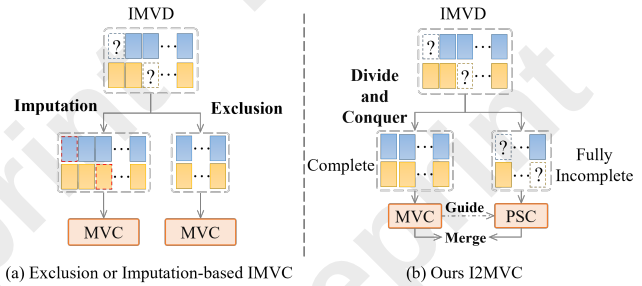


Figure 1: The difference between our `I2MVC` and other MVC methods (with two views as an example). (a) Traditional incomplete MVC methods typically convert incomplete multi-view data (IMVD) into MVD via exclusion or imputation techniques and then perform clustering using MVC methods. (b) Our imputation-free framework divides the IMVD into two parts: MVC for MVD and pseudo-supervised classification (PSC) for fully incomplete MVD.

However, multi-view data is not invariably complete. For instance, some diseases can be diagnosed without conducting all possible examinations, leading to missing views in the dataset. Such data is known as incomplete MVD. As illustrated in Figure 1 (a), a canonical approach to handling incomplete MVD involves converting it into complete MVD, followed by applying MVC methods. This conversion typically includes three strategies [Wen *et al.*, 2022]: (1) excluding data with missing views, (2) filling missing views with zero or the average of existing views [Li *et al.*, 2014; Zhao *et al.*, 2016], and (3) utilizing deep learning techniques to generate missing data [Wei *et al.*, 2020a; Bu *et al.*, 2024].

Recently, incomplete MVC solutions based on deep learning for data recovery have been put forward. For instance, COMPLETER [Lin *et al.*, 2021] recovers missing views by minimizing the conditional entropy between different views through dual prediction. ICMVC [Chao *et al.*, 2024] regards the kNN graph of existing views as a potential graph structure for the missing views and utilizes the message-passing capability of graph neural networks to fill in the missing data. The consistency and complementarity of multi-view data can be beneficial for feature learning and downstream tasks [Tan *et al.*, 2021; Wei *et al.*, 2020b]. Nevertheless, as the data scales, complex imputation becomes progressively challenging. On the one hand, interpolation of large-scale data demands a sub-

stantial amount of computing resources. On the other hand, imputation is not always effective when there is an excessive number of missing views.

Partial views can provide enough information to conduct downstream tasks [Tan *et al.*, 2018; Ren *et al.*, 2023a]. For example, the diagnosis of chronic kidney disease necessitates a comprehensive analysis of diverse data sources, including blood tests, urine tests, and kidney ultrasound. In contrast, Helicobacter pylori infection can be diagnosed solely through an antigen test. Moreover, experienced doctors can rapidly diagnose a patient's illness based on their accumulated medical knowledge without conducting additional examinations. Such rapid recognition is feasible, particularly when a certain margin of error is acceptable. Nevertheless, in the realm of incomplete multi-view clustering, using the available views to infer data categories rather than filling in the missing data through imputation is extremely challenging and remains an unsolved problem.

Based on the motivation mentioned above, this article proposes an **I**mputation-free **I**ncomplete **MVC** (`I2MVC`) algorithm. Specifically, as depicted in Figure 1 (b), the incomplete MVC problem is decomposed into a clustering problem of MVD and a classification problem of fully incomplete MVD. We train an MVC Teacher model using contrastive learning and then train a lightweight Student model for fully incomplete MVD classification through pseudo-supervised knowledge distillation. `I2MVC` brings two key benefits: (1) It fully exploits the information from MVD without requiring imputation for incomplete MVD, thereby reducing processing complexity and avoiding the accumulation of errors due to imputation. (2) Through pseudo-supervised training, the Student model can adjust the number of views in the input data to a single view, performing incomplete MVD clustering without the need of all views. Then the clustering of fully incomplete data is conducted by the Student model. Extensive experiments demonstrate the effectiveness of `I2MVC`.

The main contributions of our work are listed as follows:
(i) We propose a divide-and-conquer strategy for incomplete MVC. It converts the task into clustering complete data and pseudo-supervised classification of fully incomplete data, effectively using complete data to guide the clustering of incomplete data.
(ii) We propose an imputation-free incomplete MVC (`I2MVC`) based on pseudo-supervised knowledge distillation. `I2MVC` transforms the requirement of multiple views into a single view, reducing the required number of views when clustering incomplete data and avoiding the imputation operation.
(iii) Experiments on benchmark datasets demonstrate the effectiveness of `I2MVC` on both complete and incomplete MVD, and the pseudo-supervised knowledge distillation effectively guides the clustering of incomplete data.

2 Related Work

2.1 Multi-view Clustering

Multi-view clustering refers to methods designed to cluster complete MVD. These approaches typically include matrix factorization, multi-kernel learning, graph learning, and deep learning methods. Matrix factorization-based methods [Yao *et al.*, 2019; Wei *et al.*, 2020b; Peng *et al.*, 2023] decompose data matrices into low-rank matrices constrained by the consistency of multiple views and clustering on the low-rank matrices. Multi-kernel learning-based methods [Zhang *et al.*, 2021; Liu *et al.*, 2023; Su *et al.*, 2024] combine predefined kernels from different views, either linearly or non-linearly, to enhance clustering performance. Graph learning-based methods [Yang *et al.*, 2022] aim to learn a consistent affinity graph across all views and perform clustering algorithms like spectral clustering on the consensus graph. Deep learning-based methods [Xu *et al.*, 2022; Ren *et al.*, 2023b] leverage deep networks to learn latent representations and then use learned representations to partition the data into distinct groups.

Learning cohesive and discriminative representations from different views is crucial for deep learning-based MVC methods. Contrastive learning has been extensively employed to learn consistent representations from multiple views, presenting significant advantages in MVC. For example, MCGC [Pan and Kang, 2021] regularizes the consensus graph with a graph contrastive loss. CMHHC [Lin *et al.*, 2022] uses contrastive learning to align sample-level representations across views and performs hierarchical clustering in the hyperbolic space. DealMVC [Yang *et al.*, 2023] uses global and local contrastive calibration loss to jointly optimize interacted cross-view features at two levels for multi-view clustering. Simple Contrastive MVC (SCM) [Luo *et al.*, 2024] fuses data at the data level instead of the feature level, and constructs augmented views via masking and adding noise, training an auto-encoder with contrastive learning. Despite their strong clustering performance, these methods do not address missing views, leading to suboptimal results when applied to incomplete data.

2.2 Incomplete Multi-view Clustering

Most incomplete MVC solutions transform incomplete data into complete data through imputation and then apply MVC for clustering. Common imputation strategies include simple methods (such as zero imputation and mean imputation) [Zhou *et al.*, 2024] and generative methods (such as diffusion imputation and graph neural networks) [Wen *et al.*, 2024; Chao *et al.*, 2024]. COMPLETER [Lin *et al.*, 2021] proposes recovering missing views at sample-level, rather than missing similarity information, and achieves data recovery and consistency learning within a unified information-theoretic framework, improving the interpretability of incomplete MVC algorithms. However, sample-level recovery will lose instance commonality. ProImp [Li *et al.*, 2023] constructs a sample-prototype relationship and performs data recovery using prototypes from the missing views and the sample-prototype relationships inherited from observed views. Graph convolutional networks (GCN) [Kipf and Welling, 2017] have an advantage in learning community consistency, providing a new paradigm in imputation. ICMVC [Chao *et al.*, 2024] utilizes the message-passing mechanism of GCN to complete missing latent representations by leveraging neighbors of missing instances. DVIMC [Xu *et al.*, 2024] parametrizes the approximate posterior of each view using Variational auto-encoders (VAEs) and

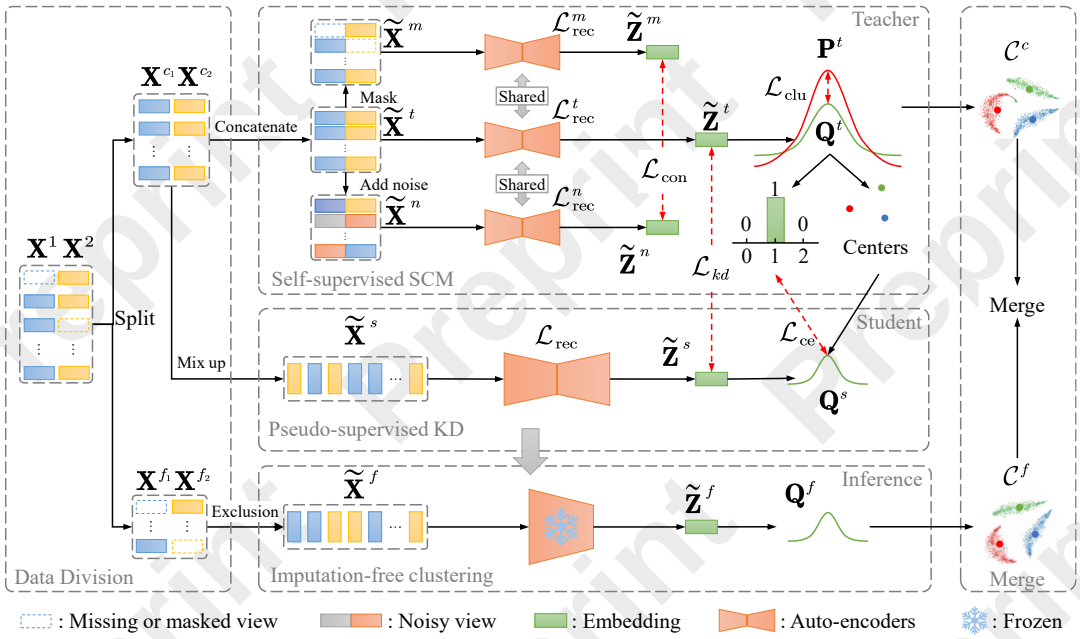


Figure 2: Conceptual framework of I2MVC. The incomplete MVD (views: V, missing rate: η) is split into two subsets based on whether views are missing: complete MVD ($\mathbf{X}^c$) and fully incomplete MVD ($\mathbf{X}^f$). For complete MVD, Self-supervised Simple Contrastive MVC is trained as the Teacher model, which provides clustering labels $\mathcal{C}^c$ and cluster centers. To free the model from the constraint of a fixed number of views, the complete MVD is mixed up to generate $(1-\eta)NV$ single-view samples. Then, the mixed data are used to train a simple Student classifier supervised by clustering pseudo-labels from the Teacher. The retained views are averaged for each sample in fully incomplete MVD, and the encoder in Student is used to infer clustering results $\mathcal{C}^f$. The final result is generated by merging $\mathcal{C}^c$ and $\mathcal{C}^f$.

integrates information through Product-of-Experts, thereby avoiding the need for imputation. Although DVIMC learns the data distribution of different views, it ignores the consistent information across views. Compared with the existing incomplete MVC methods, our I2MVC uses multi-view data for pseudo-supervised distillation and trains a Student model that can cluster incomplete data with just one view.

3 Methodology

This section introduces the proposed I2MVC, which mainly contains three components: self-supervised simple contrastive MVC, pseudo-supervised knowledge distillation, and imputation-free clustering for fully incomplete MVD. Figure 2 shows the schematic diagram of I2MVC.

3.1 Preliminary

Given an MVD dataset $\mathbf{X} = \{\mathbf{X}^1, \ldots, \mathbf{X}^v, \cdots \mathbf{X}^V\}$ with N samples and $V (V \geq 2)$ views, $\mathbf{X}^v = \{x_1^v, \cdots, x_N^v\} \in \mathbb{R}^{N \times d_v}$ is the data matrix of the v-th view, where d_v is its feature dimension. When there are no missing views in the data, it is referred as *complete MVD*. When all samples have $[1, V-1]$ missing views, it is referred as *fully incomplete MVD*. $\widetilde{\mathbf{X}}$ denotes the feature matrix formed by concatenating all views in complete MVD. For problem analysis, suppose $\widetilde{\mathbf{X}}^m$ denotes randomly masked data, and $\widetilde{\mathbf{X}}^n$ denotes the matrix with Gaussian noise. $\widetilde{\mathbf{X}}^s$ denotes mixing all views of complete MVD. For fully incomplete MVD, existing views are averaged to generate $\widetilde{\mathbf{X}}^f$. The encoder embeddings are denoted as $\widetilde{\mathbf{Z}}$, the clustering distribution as $\mathbf{Q}$, and the auxiliary distribution as $\mathbf{P}$. The clustering partition for complete MVD is $\mathcal{C}^c$, while that for fully incomplete MVD is $\mathcal{C}^f$.

3.2 Self-supervised Simple Contrastive MVC

I2MVC divides the incomplete MVD into complete MVD and fully incomplete MVD for clustering. For complete MVD, any deep MVC algorithm can be used. Simple Contrastive MVC [Luo *et al.*, 2024] efficiently transforms view-specific encoders into a single encoder via data-level fusion, thereby reducing the model size. However, its feature learning and clustering processes are isolated, causing a lack of clustering-oriented guidance during feature learning. Given that, we propose Self-supervised Simple Contrastive MVC (S3CM) to make the joint optimization of feature learning and clustering.

Before concatenation, Principal Component Analysis (PCA) [Hotelling, 1933] is applied to $\mathbf{X}$ to obtain dimension-aligned features. Consistent with Simple Contrastive MVC, data-level fusion on complete MVD is performed to obtain $\widetilde{\mathbf{X}}^t$. Then random masking and Gaussian noise are applied to generate augmented data $\widetilde{\mathbf{X}}^m$ and $\widetilde{\mathbf{X}}^n$, which are used to construct sample pairs of augmented views for contrastive learning.

Auto-encoder (AE) efficiently conducts unsupervised representation learning by minimizing the reconstruction error. Given its simplicity and effectiveness, we apply AE as a backbone network to learn the embedding of the concatenated

view and augmented views:

$$\widetilde{\mathbf{Z}}^t = f_\theta^t(\widetilde{\mathbf{X}}^t),\ \widetilde{\mathbf{Z}}^m = f_\theta^t(\widetilde{\mathbf{X}}^m),\ \widetilde{\mathbf{Z}}^n = f_\theta^t(\widetilde{\mathbf{X}}^n), \quad (1)$$

where f_θ^t is the encoder and θ is the learnable weight.

The reconstruction loss optimizes the feature learning of the AE:

$$\begin{aligned}\mathcal{L}_{rec} &= \mathcal{L}_{rec}^t + \mathcal{L}_{rec}^m + \mathcal{L}_{rec}^n,\\ \mathcal{L}_{rec}^t &= \frac{1}{N}\sum \|\hat{\mathbf{X}}^t - \widetilde{\mathbf{X}}^t\|_2^2,\\ \mathcal{L}_{rec}^m &= \frac{1}{N}\sum \|\hat{\mathbf{X}}^m - \widetilde{\mathbf{X}}^m\|_2^2,\\ \mathcal{L}_{rec}^n &= \frac{1}{N}\sum \|\hat{\mathbf{X}}^n - \widetilde{\mathbf{X}}^n\|_2^2,\end{aligned} \quad (2)$$

where $\hat{\mathbf{X}}^t, \hat{\mathbf{X}}^m, \hat{\mathbf{X}}^n$ are reconstructed data by decoder g_θ^t:

$$\hat{\mathbf{X}}^t = g_\theta^t(\widetilde{\mathbf{Z}}^t),\ \hat{\mathbf{X}}^m = g_\theta^t(\widetilde{\mathbf{Z}}^m),\ \hat{\mathbf{X}}^n = g_\theta^t(\widetilde{\mathbf{Z}}^n), \quad (3)$$

To learn the consistent information between different augmented views, a contrastive loss is constructed:

$$\mathcal{L}_{con} = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_i^{n,m} + \mathcal{L}_i^{m,n}), \quad (4)$$

where $\mathcal{L}_i^{n,m}$ and $\mathcal{L}_i^{m,n}$ are the mutual contrastive InfoNCE loss [Oord *et al.*, 2018] of the i-th instance as follows:

$$\mathcal{L}_i^{n,m} = -\log\frac{\exp(\mathbf{S}_i^{n,m}/\tau_1)}{\exp(\mathbf{S}_i^{n,m}/\tau_1) + \sum_{t\in\mathcal{S}^-}\exp(\mathbf{S}^{n,t}/\tau_1)}, \quad (5)$$

where $\mathbf{S}_i^{n,m}, \mathbf{S}^{n,t}$ are the cosine similarity between two sample features. $\mathcal{S}^- = \{\widetilde{\mathbf{z}}_j^v\}_{j\neq i}^{v=n,m}$ is the negative samples set of $\widetilde{\mathbf{z}}_i^n$. τ_1 is a temperature coefficient, which is set to 0.5 according to Simple Contrastive MVC [Luo *et al.*, 2024]. We consider different augmented views of the same sample as positive pairs and other samples as negative ones.

To ensure that representations learned by AE are aligned with the clustering objective, we use a self-supervised clustering loss to jointly optimize clustering and feature learning:

$$\begin{aligned}\mathbf{Q}_{ij} &= \frac{(1+\|\widetilde{\mathbf{z}}_i^t - \boldsymbol{\mu}_j\|_2^2)^{-1}}{\sum_{j'}(1+\|\widetilde{\mathbf{z}}_i^t - \boldsymbol{\mu}_{j'}\|_2^2)^{-1}},\\ \mathbf{P}_{ij} &= \frac{\mathbf{Q}_{ij}^2/\sum_i \mathbf{Q}_{ij}}{\sum_{j'}\mathbf{Q}_{ij'}^2/\sum_i \mathbf{Q}_{ij'}},\\ \mathcal{L}_{clu} &= \mathrm{KL}(\mathbf{P}\|\mathbf{Q}),\end{aligned} \quad (6)$$

where $\boldsymbol{\mu}_j$ is the cluster centers vector of the j-th cluster, and $\mathbf{Q}_{ij}$ is the probability of the i-th instance assigned to the j-th cluster. $\mathbf{P}$ is a target distribution for promoting high-confident probabilities to be larger and low-confident ones to be smaller, which is controlled by the KL divergence [Kullback and Leibler, 1951]. To obtain reliable initial clustering centers for calculating clustering distribution $\mathbf{Q}$, we pre-train AE $f_\theta^t,\ g_\theta^t$ using the loss function as follows:

$$\mathcal{L}_{pre}^t = \mathcal{L}_{rec} + \mathcal{L}_{con}. \quad (7)$$

The initial clustering centers are initialized by conducting k-means [Hartigan and Wong, 1979] on the embedding learned by AE. The S3CM is optimized by $\mathcal{L}^t$:

$$\mathcal{L}^t = \mathcal{L}_{con} + \alpha\cdot\mathcal{L}_{clu}, \quad (8)$$

where α is a hyper-parameter controlling self-supervised clustering.

Finally, the clustering labels of complete MVD are obtained by selecting the index with maximum probability:

$$\mathcal{C}_i^c = \arg\max_j \mathbf{Q}_{ij},\ j = 1,\cdots,\ k, \quad (9)$$

where k is the pre-defined number of clusters.

3.3 Pseudo-supervised Knowledge Distillation

S3CM uses contrastive loss, reconstruction loss, and self-supervised clustering loss to fully learn information from complete MVD, which can be applied for feature learning and clustering in fully incomplete MVD. However, due to the missing views, its formulation cannot directly fit the input requirement of S3CM.

To enable clustering using only the available views in fully incomplete MVD, we transfer the knowledge from S3CM to a simpler AE through knowledge distillation [Hinton, 2015]. This involves training a model capable of learning single-view features. Specifically, we first adapt the AE's input by mixing up the views in complete MVD to generate $\widetilde{\mathbf{X}}^s$, where each view represents a separate sample. The AE then learns data features through the following process:

$$\widetilde{\mathbf{Z}}^s = f_\theta^s(\widetilde{\mathbf{X}}^s),\ \hat{\mathbf{X}}^s = g_\theta^s(\widetilde{\mathbf{Z}}^s). \quad (10)$$

It is also optimized by reconstruction loss $\mathcal{L}_{rec}^s$. The soft clustering assignment $\mathbf{Q}^s$ is calculated according to the well-trained center vectors. Additionally, to enable the Student model to learn features without ground-truth supervision, we design the following offline distillation process:

$$\begin{aligned}\mathcal{L}_{kd} &= \mathrm{KL}(\mathrm{softmax}(\widetilde{\mathbf{Z}}^t/\tau_2)\|\mathrm{softmax}(\widetilde{\mathbf{Z}}^s/\tau_2)),\\ \mathcal{L}_{ce} &= \mathrm{CrossEntropy}(\mathbf{Q}^s, \mathrm{OneHot}(\mathcal{C}^c)),\\ \mathcal{L}^s &= \mathcal{L}_{rec}^s + (1-\beta)\cdot\mathcal{L}_{kd} + \beta\cdot\mathcal{L}_{ce},\end{aligned} \quad (11)$$

where τ_2 is the distillation temperature and β is also a balanced hyper-parameter. Since no ground-truth labels exist, the one-hot vector of clustering pseudo-labels is substituted for the classification pseudo-supervision information. This ensures that the Student AE can generalize the knowledge learned by S3CM, enabling feature extraction and clustering for fully incomplete MVD with its observed views.

3.4 Imputation-free Clustering for Fully Incomplete MVD

Considering that the number of available views in fully incomplete MVD varies across samples, we compute the feature $\widetilde{\mathbf{X}}^f$ for each sample by averaging the features of its available views. The Student model then clusters these samples, producing the clustering result $\mathcal{C}^f$ for fully incomplete MVD:

$$\begin{aligned}\widetilde{\mathbf{Z}}^f &= f_\theta^s(\widetilde{\mathbf{X}}^f),\\ \mathbf{Q}_{ij}^f &= \frac{(1+\|\widetilde{\mathbf{z}}_i^f - \boldsymbol{\mu}_j\|_2^2)^{-1}}{\sum_{j'}(1+\|\widetilde{\mathbf{z}}_i^f - \boldsymbol{\mu}_{j'}\|)^{-1}},\\ \mathcal{C}^f &= \arg\max_j \mathbf{Q}_{ij}^f,\ j = 1,\cdots,\ k.\end{aligned} \quad (12)$$

Finally, the clustering partitions $\mathcal{C}^c$, $\mathcal{C}^f$ from complete MVD and fully incomplete MVD are simply merged to obtain the final clustering partition $\mathcal{C}$, since $\mathcal{C}^c$, $\mathcal{C}^f$ are induced from the same cluster centers.

3.5 Analysis of Computational Complexity

Assume the incomplete dataset contains N samples, V views, and k clusters, with a view missing rate of η. `I2MVC` adopts mini-batch training, with the batch size denoted as $|\mathcal{B}|$. The computational complexity of each stage is analyzed as follows: In the data division stage, the algorithm determines whether each sample has complete views and partitions the data accordingly. This operation has a complexity of $\mathcal{O}(NV)$. The complexity of the S3CM phase primarily arises from the contrastive loss computation, resulting in a complexity of $\mathcal{O}(((1-\eta)N)^2)$. During the pseudo-supervised knowledge distillation phase, the complexity is $\mathcal{O}((1-\eta)Nk)$, as it involves assigning pseudo-labels and training the lightweight classifier on $(1-\eta)N$ complete MVD samples. The imputation-free clustering phase, responsible for clustering ηN fully incomplete MVD samples, has a complexity of $\mathcal{O}(\eta Nk)$. For incomplete data clustering, overall complexity is $\mathcal{O}(|\mathcal{B}|^2 + NV + |\mathcal{B}|k)$; for complete data clustering, it is $\mathcal{O}(|\mathcal{B}|^2 + |\mathcal{B}|k)$, both dominated by S3CM.

4 Experiments

4.1 Experimental Setup

Datasets. The datasets used in this study are benchmark datasets widely used in MVC, including NGs [Hussain *et al.*, 2010], BDGP [Cai *et al.*, 2012], WebKB [Craven *et al.*, 1998], and MNIST-USPS [Peng *et al.*, 2019]. These datasets not only contain different modalities and different amounts of samples, but also different numbers of views. Table 1 provides the statistics information of these datasets.

Since our focus is on incomplete MVD, we randomly selected data from each category to construct two subsets: complete MVD and fully incomplete MVD, with a missing rate of η, where $0 \leq \eta < 1$. To create the fully incomplete MVD, we applied random masking to set the number of views for each sample to a range of $[1, V-1]$, where V is the total number of views. This setup ensures that the constructed datasets simulate scenarios involving randomly missing views.

Metrics. The performance of all methods was assessed using three primary evaluation metrics: clustering accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (ARI). To enable comprehensive algorithm comparison, we additionally introduced four computational efficiency indicators: peak GPU memory utilization (GPU MEM), total trainable parameters (PARAM), computational time per epoch (TPE), and total execution time (TIME). These metrics collectively serve to quantify both the effectiveness and efficiency of the evaluated approaches.

Baselines. Baselines span three categories: complete MVC (DIVIDE [Zhang *et al.*, 2024] and SCM [Luo *et al.*, 2024]), imputation-based incomplete MVC (COMPELTER [Lin *et al.*, 2021], ProImp [Li *et al.*, 2023], ICMVC [Chao *et al.*, 2024]), and imputation-free incomplete MVC (APADC [Xu *et al.*, 2023] and GIMVC [Bai *et al.*, 2024]). All baselines conduct clustering using k-means [Hartigan and Wong, 1979] on the latent representation of data.

Dataset	NGs	WebKB	MNIST-USPS	BDGP
Modality	Text	Image	Image	Text&Image
#Samples	500	1051	5000	2500
#Classes	5	2	10	5
#Views	3	2	2	2
#Dimension	2000/2000/2000	1840/3000	784/784	1750/79

Table 1: Statistics of four benchmark datasets.

Implementation Details. We used the publicly available code for the compared methods and followed the implementation details provided in their original papers. For `I2MVC`, its backbone network structure was kept consistent with SCM [Luo *et al.*, 2024] to ensure a fair comparison. The code of `I2MVC` was developed using Python 3.8 and PyTorch 1.13, and all experiments were conducted on a Tesla T4 GPU with 16GB memory. The distillation temperature τ_2 was set to 2, and the view missing rate varied within $\{0,\ 0.1,\ 0.3,\ 0.5,\ 0.7\}$. The learning rate for the Teacher model was set to 3e-4 during both the pre-training and training phases, while the Student model in the distillation phase used a learning rate of 3e-5. Each phase was trained for at least 50 iterations. The hyper-parameters $\alpha,\ \beta$ were searched within $\{0.001,\ 0.01,\ 0.1,\ 1,\ 10\}$ and ranged $0.1 \sim 0.9$ with step 0.1, respectively. All algorithms were performed five times, and the average value and standard deviation were measured.

4.2 Comparison with Baselines on Complete and Incomplete MVD

We compared the clustering performance of eight methods across four datasets, including both complete MVD and incomplete MVD (with a missing rate of η=50%). The results are shown in Table 2. Based on the results, we make the following observations: (i) Our proposed `I2MVC` achieves competitive results on both complete and incomplete MVD, especially on incomplete data. (ii) Although complete MVC algorithms are impacted by view missing, applying simple filling techniques can still yield reasonable clustering performance, which suggests it is possible to achieve MVC without specific imputation. (iii) Imputation-based clustering methods generally outperform other imputation-free methods except for our `I2MVC`. Intuitively, employing imputation to complement missing views will introduce additional information. However, when the view missing rate is high, the quality of the supplemented views is compromised, leading to suboptimal clustering outcomes. (iv) Since our `I2MVC` uses the clustering partitions of the complete view data as supervision information to guide the clustering of incomplete data, `I2MVC` stands out among the imputation-free methods. These results demonstrate the effectiveness of `I2MVC`.

4.3 Performance with Different Missing Rates

Figure 3 presents the clustering performance of eight methods applied to the BDGP with varying missing rates within $\{0,\ 0.1,\ 0.3,\ 0.5,\ 0.7\}$. `I2MVC` exhibits a more pronounced

η	Method	NGs			BDGP			MNIST-USPS			WebKB		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0%	DIVIDE	43.6±4.9	21.9±1.5	18.5±2.3	89.2±5.5	78.6±7.5	76.1±10.7	94.3±5.3	92.3±1.2	91.2±4.3	89.2±5.5	**78.6±7.5**	76.1±10.7
	SCM	96.3±0.9	89.0±2.1	91.0±2.1	95.0±2.4	87.1±2.8	88.4±4.7	98.7±0.0	96.5±0.1	97.2±0.0	67.6±4.9	23.4±3.0	12.7±6.5
	COMP	39.1±6.0	16.0±5.9	9.1±3.8	60.2±2.8	55.8±3.8	34.3±5.1	87.1±7.6	93.6±2.5	86.5±7.0	85.0±5.3	25.6±19.1	35.0±24.5
	ProImp	74.5±11.9	55.3±9.7	52.5±12.0	71.3±13.7	64.9±15.9	54.3±21.0	**99.5±0.1**	**98.7±0.2**	**99.0±0.2**	75.6±8.3	26.6±15.6	28.2±17.8
	ICMVC	87.7±2.6	68.5±4.3	72.0±4.9	93.4±9.3	89.4±9.3	88.9±12.7	99.3±0.1	98.0±0.4	98.5±0.3	73.8±1.5	30.2±4.5	22.7±2.7
	APADC	40.1±5.3	14.4±5.3	11.4±5.3	66.0±5.6	50.1±2.8	40.2±5.9	97.6±0.6	94.7±0.9	94.6±1.5	84.4±3.5	25.4±8.0	40.2±9.2
	GIMVC	61.2±1.1	51.0±2.0	37.4±2.0	91.9±0.0	77.7±0.1	80.9±0.1	87.1±6.0	94.6±2.1	81.1±5.1	89.9±0.3	41.0±1.4	58.1±1.0
	I2MVC	**97.7±0.5**	**92.8±1.3**	**94.3±1.1**	**98.1±0.5**	**94.5±1.0**	**95.5±1.1**	98.7±0.1	96.5±0.2	97.2±0.2	**97.4±0.5**	78.5±3.1	**88.8±1.9**
50%	DIVIDE	29.2±3.0	7.0±3.9	5.1±3.7	55.6±4.5	46.5±2.1	30.1±4.3	91.9±0.7	83.1±0.8	83.1±1.2	76.9±8.1	**57.0±8.1**	53.1±11.3
	SCM	65.8±2.9	52.7±3.3	47.0±3.9	85.5±1.0	66.4±1.6	67.0±2.3	92.2±0.2	84.4±0.3	83.8±0.3	65.5±3.6	22.4±4.8	9.5±4.9
	COMP	32.4±3.5	7.4±2.2	4.0±2.0	57.2±5.5	43.9±5.4	25.9±3.5	87.3±0.2	89.3±0.5	83.9±0.5	78.8±2.9	5.6±6.8	9.5±13.5
	ProImp	55.4±6.2	31.5±5.1	27.7±6.1	58.9±7.3	45.9±7.6	37.4±8.7	96.7±0.4	91.8±0.8	92.8±0.8	70.4±4.0	16.0±4.7	16.8±5.9
	ICMVC	54.0±6.1	27.9±4.2	24.9±4.6	84.8±5.1	71.1±4.5	70.0±6.4	96.0±0.1	90.5±0.1	91.4±0.1	68.5±1.6	15.2±2.5	13.7±2.3
	APADC	26.3±2.0	4.4±1.7	1.8±1.5	55.9±4.1	32.9±5.4	28.7±6.2	94.8±0.6	88.6±0.8	88.5±1.4	78.1±0.3	0.9±1.2	1.7±3.2
	GIMVC	37.5±1.9	25.9±0.9	11.9±0.5	87.2±0.5	68.8±0.6	70.9±1.0	81.5±3.1	78.5±0.7	73.0±1.4	85.4±0.3	27.2±1.3	44.4±1.4
	I2MVC	**92.0±0.8**	**78.5±1.5**	**80.9±1.7**	**96.8±1.1**	**90.4±2.4**	**92.1±2.6**	**96.8±0.1**	**92.0±0.3**	**93.0±0.3**	**91.8±0.3**	50.8±1.9	**64.6±1.1**

Table 2: Clustering results of all algorithms on four datasets with different missing rates (0% means complete data and 50% means incomplete data). The best is highlighted in bold, and the underline denotes the second best.

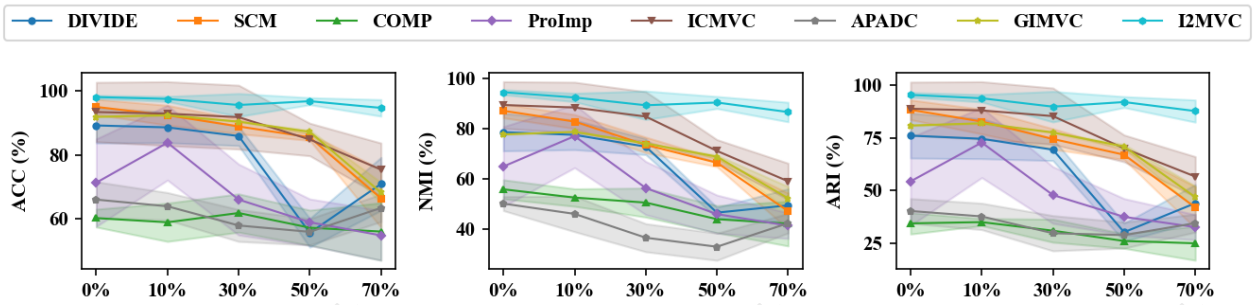


Figure 3: Clustering performance comparison of eight methods on BDGP with different missing rates.

clustering effect on data with low missing rates than other methods. Moreover, the proposed approach retains its ability to effectively cluster data despite the increase of the missing rate. These findings demonstrate the effectiveness of the proposed clustering framework.

4.4 Ablation Study

To validate the effectiveness of the key components (i.e. S3CM and pseudo-supervised knowledge distillation) in our proposed method, we designed several variant models: (i) **SCM** is the simple contrastive multi-view clustering network [Luo *et al.*, 2024]; (ii) `I2MVC` **w/o S3CM** removes the proposed S3CM module and directly uses the MVD information learned by original SCM to guide the training of the Student network; (iii) `I2MVC` **w/o PSKD** removes the pseudo-supervised knowledge distillation (PSKD) module from the `I2MVC` framework. In this case, the Student model is trained solely using the reconstruction loss of the auto-encoder, without guidance from the well-trained Teacher model on the MVD. Besides, we recorded the clustering performance at each clustering stage.

The ablation experiments were conducted using the NGs and BDGP datasets, with a missing rate $\eta = 0.5$. The results about different components are presented in Figure 4. Overall, removing any single component from `I2MVC` leads to a

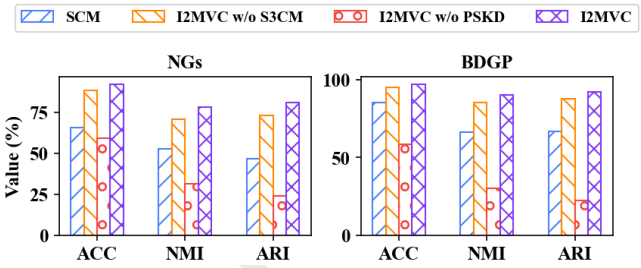


Figure 4: Ablation results of different components.

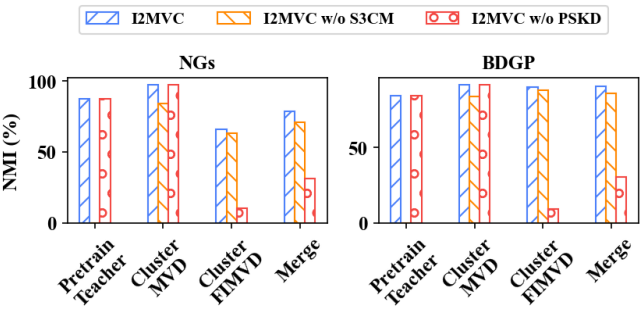
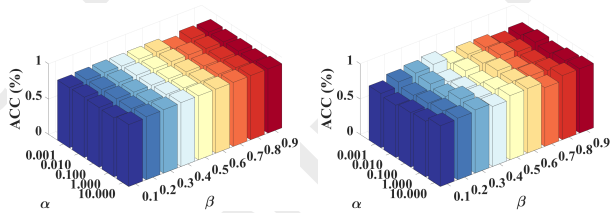


Figure 5: Ablation results of different clustering stages.



(a) NGs (b) BDGP

Figure 6: Clustering results with different hyper-parameters.

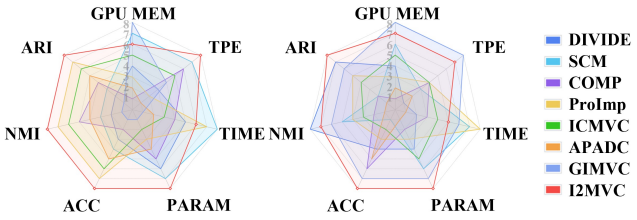


Figure 7: Comprehensive performance comparison of eight methods on BDGP and MNIST-USPS.

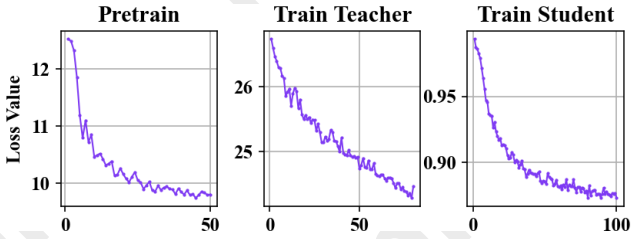


Figure 8: Loss value at different training stages on BDGP.

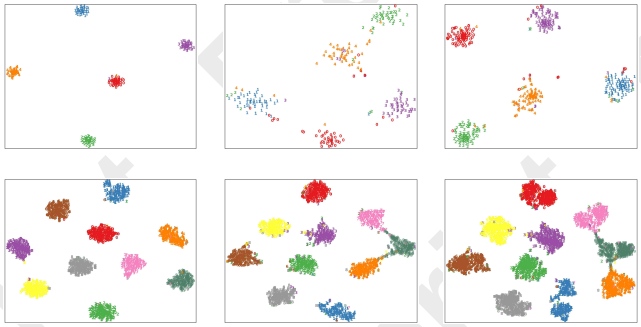


Figure 9: Visualization of clustering results with missing rate 50%. From left to right: complete MVD, fully incomplete MVD, and Merged clusters. From top to bottom: NGs and MNIST-USPS.

performance decrease, which indicates that each component of the proposed method is effective. Specifically, after removing the self-supervised clustering loss, the clustering performance on both datasets significantly deteriorates, demonstrating the important role of the proposed S3CM in learning clustering-friendly representations. Additionally, after removing the pseudo-supervised knowledge distillation, the clustering performance on both datasets decreases sharply, suggesting that the hard clustering pseudo-labels and soft clustering logits generated by S3CM play a guiding role when training the Student model. As illustrated in Figure 5, the results across distinct clustering stages reveal that S3CM not only significantly enhances the clustering performance of the Teacher module but also exerts a notable influence on clustering outcomes in subsequent phases. In contrast, PSKD primarily affects the clustering efficacy of fully incomplete MVD (FIMVD).

4.5 Model Analyses

Hyper-parameters. `I2MVC` involves hyper-parameters during training: α that controls the self-supervised clustering in S3CM, and β that controls pseudo-supervised knowledge distillation. We conducted a hyper-parameter search on the NGs and BDGP datasets, with the results visualized in a 3D bar chart in Figure 6. The results indicate that `I2MVC` is not very sensitive to these hyper-parameters. Generally, the value of α can be set to 1, and β can be set to 0.5.

Comprehensive Performance. Figure 7 presents a comparative analysis of the comprehensive performance across competing methods. The area occupied by each method within the radar chart is proportional to its aggregated performance. Evidently, `I2MVC` gains significant superiority in integrated performance compared to existing MVC and IMVC baselines.

Convergence. Figure 8 plots the loss function values across distinct clustering stages on the BDGP dataset. The results demonstrate that the proposed `I2MVC` exhibits a consistent loss decline across all stages, ultimately achieving robust convergence.

Visualization. We used t-SNE [van der Maaten and Hinton, 2008] to visualize the clustering results and features with 50% missing views to intuitively demonstrate the clustering performance of `I2MVC` on both complete MVD and incomplete MVD. As shown in Figure 9, `I2MVC` not only clusters complete data effectively but also learns information from complete data to guide the clustering of incomplete data. Moreover, the clustering results of incomplete data indicate that it is feasible to achieve effective clustering of incomplete data using only the available views without imputation.

5 Conclusion

We propose an imputation-free incomplete MVC framework (`I2MVC`) via pseudo-supervised knowledge distillation, to address the error accumulation issue inherent in existing imputation-based MVC methods. Experimental results show that `I2MVC` not only achieves superior clustering performance on incomplete MVD but also performs well on complete MVD. Future work will focus on exploring more flexible mix-up strategies.

Acknowledgements

This work is supported by National Key Research and Development Program of China (No. 2024YFF1206604), NSFC

(62272276 and 62432006), Shandong Provincial Natural Science Foundation (No. ZR2024JQ001), Taishan Scholars Program (No. tsqn202306007 and tsqn202408317).

References

[Bai *et al.*, 2024] Shunshun Bai, Qinghai Zheng, Xiaojin Ren, and Jihua Zhu. Graph-guided imputation-free incomplete multi-view clustering. *ESWA*, 258:125165, 2024.

[Bickel and Scheffer, 2004] Steffen Bickel and Tobias Scheffer. Multi-view clustering. In *ICDM*, pages 19–26, 2004.

[Bu *et al.*, 2024] Yongqi Bu, Jiaxuan Liang, Zhen Li, Jianbo Wang, Jun Wang, and Guoxian Yu. Cancer molecular subtyping using limited multi-omics data with missingness. *PLOS Comp. Biol.*, 20(12):e1012710, 2024.

[Cai *et al.*, 2012] Xiao Cai, Hua Wang, Heng Huang, and Chris Ding. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinfo.*, 28(12):i16–i24, 2012.

[Chao *et al.*, 2024] Guoqing Chao, Yi Jiang, and Dianhui Chu. Incomplete contrastive multi-view clustering with high-confidence guiding. In *AAAI*, pages 11221–11229, 2024.

[Craven *et al.*, 1998] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Seán Slattery. Learning to extract symbolic knowledge from the world wide web. In *AAAI*, pages 509–516, 1998.

[Hartigan and Wong, 1979] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *J R Stat Soc C-Appl*, 28(1):100–108, 1979.

[Hinton, 2015] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Hotelling, 1933] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *J Educ. Psychol.*, 24(6):417, 1933.

[Hussain *et al.*, 2010] Syed Fawad Hussain, Gilles Bisson, and Clément Grimal. An improved co-similarity measure for document clustering. In *ICMLA*, pages 190–197, 2010.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[Kullback and Leibler, 1951] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Math. Stat.*, 22(1):79–86, 1951.

[Li *et al.*, 2014] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *AAAI*, pages 1968–1974, 2014.

[Li *et al.*, 2023] Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. Incomplete multi-view clustering via prototype-based imputation. In *IJCAI*, pages 3911–3919, 2023.

[Lin *et al.*, 2021] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *CVPR*, pages 11174–11183, 2021.

[Lin *et al.*, 2022] Fangfei Lin, Bing Bai, Kun Bai, Yazhou Ren, Peng Zhao, and Zenglin Xu. Contrastive multi-view hyperbolic hierarchical clustering. In *IJCAI*, pages 3250–3256, 2022.

[Liu *et al.*, 2023] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Qing Liao, and Yuanqing Xia. Contrastive multi-view kernel learning. *TPAMI*, 45(8):9552–9566, 2023.

[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 *IJCAI*, pages 4697–4705, 2024.

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

[Pan and Kang, 2021] Erlin Pan and Zhao Kang. Multi-view contrastive graph clustering. In *NeurIPS*, pages 2148–2159, 2021.

[Peng *et al.*, 2019] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *ICML*, pages 5092–5101, 2019.

[Peng *et al.*, 2023] Siyuan Peng, Jingxing Yin, Zhijing Yang, Badong Chen, and Zhiping Lin. Multiview clustering via hypergraph induced semi-supervised symmetric nonnegative matrix factorization. *TCVST*, 33(10):5510–5524, 2023.

[Ren *et al.*, 2023a] Liangrui Ren, Jun Wang, Zhao Li, Qingzhong Li, and Guoxian Yu. scmcs: a framework for single-cell multi-omics data integration and multiple clusterings. *Bioinf.*, 39(4):btad133, 2023.

[Ren *et al.*, 2023b] Liangrui Ren, Guoxian Yu, Jun Wang, Lei Liu, Carlotta Domeniconi, and Xiangliang Zhang. A diversified attention model for interpretable multiple clusterings. *TKDE*, 35(9):8852–8864, 2023.

[Su *et al.*, 2024] Peng Su, Yixi Liu, Shujian Li, Shudong Huang, and Jiancheng Lv. Robust contrastive multi-view kernel clustering. In *IJCAI*, pages 4938–4945, 2024.

[Tan *et al.*, 2018] Qiaoyu Tan, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Incomplete multi-view weak-label learning. In *IJCAI*, pages 2703–2709, 2018.

[Tan *et al.*, 2021] Qiaoyu Tan, Guoxian Yu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. Individuality- and commonality-based multiview multilabel learning. *TCYB*, 51(3):1716–1727, 2021.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

[Wei *et al.*, 2020a] Shaowei Wei, Jun Wang, Guoxian Yu, Carlotta Domeniconi, and Xiangliang Zhang. Deep incomplete multi-view multiple clusterings. In *ICDM*, pages 651–660, 2020.

[Wei *et al.*, 2020b] Shaowei Wei, Jun Wang, Guoxian Yu, Carlotta Domeniconi, and Xiangliang Zhang. Multi-view multiple clusterings using deep matrix factorization. In *AAAI*, pages 6348–6355, 2020.

[Wei *et al.*, 2021] Shaowei Wei, Guoxian Yu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. Multiple clusterings of heterogeneous information networks. *Mach. Learn.*, 110(6):1505–1526, 2021.

[Wen *et al.*, 2022] Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang, and Jinxing Li. A survey on incomplete multiview clustering. *TSMCS*, 53(2):1136–1149, 2022.

[Wen *et al.*, 2024] Jie Wen, Shijie Deng, Waikeung Wong, Guoqing Chao, Chao Huang, Lunke Fei, and Yong Xu. Diffusion-based missing-view generation with the application on incomplete multi-view clustering. In *ICML*, pages 52762–52778, 2024.

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

[Xu *et al.*, 2023] Jie Xu, Chao Li, Liang Peng, Yazhou Ren, Xiaoshuang Shi, Heng Tao Shen, and Xiaofeng Zhu. Adaptive feature projection with distribution alignment for deep incomplete multi-view clustering. *TIP*, 32:1354–1366, 2023.

[Xu *et al.*, 2024] Gehui Xu, Jie Wen, Chengliang Liu, Bing Hu, Yicheng Liu, Lunke Fei, and Wei Wang. Deep variational incomplete multi-view clustering: Exploring shared clustering structures. In *AAAI*, pages 16147–16155, 2024.

[Yang *et al.*, 2022] Ben Yang, Xuetao Zhang, Feiping Nie, and Fei Wang. Fast multiview clustering with spectral embedding. *TIP*, 31:3884–3895, 2022.

[Yang *et al.*, 2023] Xihong Yang, Jin Jiaqi, 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 *ACM MM*, page 337–346, 2023.

[Yao *et al.*, 2019] Shixin Yao, Guoxian Yu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. Multi-view multiple clustering. In *IJCAI*, pages 4121–4127, 2019.

[Yu *et al.*, 2024] Guoxian Yu, Liangrui Ren, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. Multiple clusterings: Recent advances and perspectives. *Comp. Sci. Rev.*, 52:100621, 2024.

[Zhang *et al.*, 2021] Tiejian Zhang, Xinwang Liu, Lei Gong, Siwei Wang, Xin Niu, and Li Shen. Late fusion multiple kernel clustering with local kernel alignment maximization. *TMM*, 25:993–1007, 2021.

[Zhang *et al.*, 2024] Chao Zhang, Deng Xu, Xiuyi Jia, Chunlin Chen, and Huaxiong Li. Continual multi-view clustering with consistent anchor guidance. In *IJCAI*, pages 5434–5442, 2024.

[Zhao *et al.*, 2016] Handong Zhao, Hongfu Liu, and Yun Fu. Incomplete multi-modal visual data grouping. In *IJCAI*, pages 2392–2398, 2016.

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