Deep Safe Multi-view Clustering: Reducing the Risk of Clustering Performance Degradation Caused by View Increase

Huayi Tang[1,2], Yong Liu[1,2] *
[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China
tangh4681@gmail.com, liuyonggsai@ruc.edu.cn

Abstract

Multi-view clustering has been shown to boost clustering performance by effectively mining the complementary information from multiple views. However, we observe that sometimes learning from data with more views is not guaranteed to achieve better clustering performance than from data with fewer views. To address this issue, we propose a general deep learning based framework which is guaranteed to reduce the risk of performance degradation caused by view increase. Concretely, the model is required to extract complementary information and discard the meaningless noise by automatically selecting features. These two learning procedures are incorporated into one unified framework by the proposed optimization objective. In theory, the empirical clustering risk of the proposed framework is no higher than learning from data before the view increase and data of the new increased single view. Also, the expected clustering risk of the framework under divergence-based loss is no higher than that with high probability. Comprehensive experiments on benchmark datasets demonstrate the effectiveness and superiority of the proposed framework in achieving safe multi-view clustering.

1. Introduction

Multi-view data, which contains data collected from different sources, exists widely in real-world application scenarios. For example, video can be represented by audible and visual information, and images can be characterized by different descriptors. As an important topic in multi-view learning, multi-view clustering (MVC) aims to partition similar instances into the same group and dissimilar instances into different groups by utilizing the complementary information in multi-view data [36, 44]. Through the well-designed learning mechanism, multi-view clustering can fully discover the potential structure hidden in multi-view data and achieve better clustering performance.

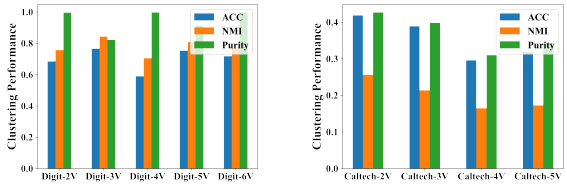


Figure 1. Clustering performance degradation phenomenon on Digit and Caltech datasets.

However, in real-world applications, multi-view data are collected and constructed dynamically, which leads to an increasing number of views. For instance, a new view will be added to the original multi-view dataset after a new description is proposed. Thus, a direct problem is, will the clustering performance of the multi-view model degrade when the number of views increases? Intuitively, data of the increased view contain both semantic features and meaningless noise. The former can provide complementary information that is beneficial for improving clustering performance. The latter, however, may bring the risk of clustering performance degradation. That is, more views do not necessarily guarantee to promote the clustering performance. Sometimes, on the contrary, conducting clustering on the dataset with more views may obtain worse results than that with fewer views. This performance degradation caused by view increase is observed in our experiments. As shown in Figure 1, the clustering performance of some MVC methods degenerates when the number of views increases, which verifies the fact that conducting clustering on datasets with more views will not always be better. Besides, single-view can be regarded as a special variant of multi-view. Thus, any multi-view model can be directly applied to obtain a single-view result on data of the new increased view. This result may perform better than that of the multi-view model, which has been verified and discussed in [35]. Therefore,

*Corresponding author.

how to reduce the risk of clustering performance degradation caused by view increase should be considered from both single-view and multi-view aspects.

Although many solid MVC methods [26,29,31,41] have been proposed, the efforts to tackle the clustering performance degradation caused by view increase are still limited. To this end, we aims to design an theoretical framework that provides a lower bound performance guarantee for MVC methods, such that more views never hurt the clustering performance. In this paper, we firstly give a formal and complete definition of *safe multi-view clustering*. However, the main challenge of achieving safe multi-view clustering is that the ground-truth labels are not available. Thus, from the perspectives of empirical and expected clustering risk, we introduce the definitions of empirical and $(\epsilon, \delta, \delta_n)$-expected safe multi-view clustering. Based on that, we propose a general deep learning based multi-view clustering framework. From the perspective of clustering and representation learning, the model is required to extract complementary information from multi-view data. Meanwhile, from the perspective of safeness, the model is also required to automatically select the features from single view or multiple views. That is, if data of the new increased view contain more meaningless noise than useful complementary information, the features learned from this view should be discarded. These two learning processes are cast as a unified optimization problem. In theory, the proposed framework is guaranteed to achieve empirical safe multi-view clustering. Also, we discuss a special case of the proposed framework under divergence-based loss and prove that it can achieve the defined $(\epsilon, \delta, \delta_n)$-expected safe multi-view clustering. Experiments on benchmark datasets demonstrate the effectiveness of the proposed learning mechanism to achieve safe multi-view clustering.

2. Related Work

In this section, we briefly introduce the most related work to our study in this paper, including multi-view clustering and safeness studies in machine learning.

Multi-view Clustering. Existing multi-view clustering methods can be divided into five categories, including multiple kernel learning based approaches [23,25–27,38], spectral based approaches [11, 15, 45, 46], subspace learning based approaches [1, 13, 34, 41], non-negative matrix factorization based approaches [2, 24, 42], and deep learning based approaches [36, 39, 40, 44]. In [38], the authors propose a late-fusion method where the weighted basic partitions are aligned to obtain a consensus partition. Zhang *et al.* [43] propose a unified framework where binary representation learning and binary clustering are jointly conducted. In [17], view-peculiar subspace representations are mined and integrated into a common latent representation to extract complementary and census information from multiple views. The work in [31] conducts multi-view clustering via connection graph to achieve geometric consistency and cluster assignment consistency. In [32], a self-pace learning mechanism is introduced to address the issue that being stuck in local optima. Zhou *et al.* [44] propose an end-to-end clustering framework where the latent features are mined by adversarial learning as well as attention mechanism. In [36], contrastive learning is adopted to prevent the model from learning a group of equal fusion weights, which is demonstrated to obtain more accurate results. Liu *et al.* [26] propose a late-fusion framework that combines the cluster assignments generation and the learning of consensus partition matrix into an unified procedure. The work in [27] calculates the kernel alignment in a local manner. Recent MVC methods [39, 40] enjoy good representation and clustering capabilities on benchmark datasets. Recent works have focused on the robustness of multi-view learning. In [29], a new weight learning schema is proposed to learn proper weights for multiple views. The authors in [10] design a new multi-view classification framework to promote the classification reliability via integrating multiple views at an evidence level. Unlike the above methods, we propose a new framework which aims to reduce the potential risk of clustering performance caused by view increase and provide a guarantee of no worse clustering performance in the cases where views are dynamically increase.

Safeness Studies in Machine Learning. Safeness is an important topic in machine learning which focus on reducing the performance degradation of learners. There are some pioneer works that achieve safeness in semi-supervised learning and weakly supervised learning [8, 19–21]. Li *et al.* [21] propose safe semi-supervised support vector machines by utilizing low-density separators. In [19], safe predictions are learned from multiple semi-supervised regressors by the proposed projection algorithm. A general ensemble learning schema that integrates multiple weakly supervised learners is presented in [20]. Recently, much work have focused on achieving the safeness of deep learning methods. The authors in [8] provide a novel safe deep semi-supervised learning framework to address the performance degradation caused by class distribution mismatch. These methods focus on classification and regression tasks on single-view data and not suitable for multi-view data without ground-truth labels. [35] is the first work to achieve safeness in multi-view clustering, which obtains safe cluster assignments that are no worse than given single-view methods based on several candidate multi-view clustering by solving a max-min optimization problem. Moreover, a safe multi-task model is proposed in [9] which is guaranteed to be no worse than its single-task component on each task. Different from the above methods, our work is to guarantee that the new increasing view never degrades the clustering performance on the previous views.

3. The Proposed Method

In this section, we first give the notations used in this paper. Then several definitions of safe multi-view clustering are introduced, including safe multi-view clustering, empirical safe multi-view clustering, and $(\epsilon, \delta, \delta_n)$-expected safe multi-view clustering. After that, we present a general deep learning based multi-view clustering framework and theoretically demonstrate its mechanism to achieve safeness.

3.1. Notations

$\mathcal{D} = \{\mathbf{x}_i^1, \cdots, \mathbf{x}_i^m\}_{i=1}^n$ denotes multi-view data with m views sampled i.i.d from a certain distribution μ over input space $\mathcal{X}$. K is the number of clusters. $\mathbf{A}_{i,:}$ and $\mathbf{A}_{:,j}$ denote the i-th row and j-th column of matrix $\mathbf{A}$, respectively. $\binom{K}{2}$ denotes the combination number. $\hat{\mathcal{L}}_n^{\{1,\cdots,p-1\}}$ and $\hat{\mathcal{L}}_n^p$ denote the empirical clustering risk of the multi-view model learning from data before view increase and data of the new increased view, respectively. $\hat{\mathcal{L}}_n$ denotes the empirical clustering risk of the safe multi-view model. The expectation of $\hat{\mathcal{L}}^{\{1,\cdots,p-1\}}$, $\hat{\mathcal{L}}_n^p$ and $\hat{\mathcal{L}}_n$ are denoted as $\mathcal{L}^{\{1,\cdots,p-1\}}$, $\mathcal{L}^p$, and $\mathcal{L}$, respectively.

3.2. Definitions

When the ground-truth labels are available, one can measure the ability of the model to achieve safeness (*i.e.*, performing no worse than the model that learns from data before view increase and the data of the new increased view) by comparing the clustering results with the ground-truth labels, which naturally leads to the following definition.

Definition 1 (Safe Multi-view Clustering). *If the clustering performance of a multi-view model is no worse than learning from data before view increase and the data of the increased view when the number of views increases, this model is said to achieve safe multi-view clustering.*

However, the clustering performance of the model is unknown during the learning process as the ground-truth labels are not available, which makes it hard to evaluate the ability of the model to achieve safeness by Definition 1. According to the empirical risk minimization, the model should minimize the empirical clustering risk on multi-view data. With this observation in mind, we presents the following definition to describe the empirical safeness of MVC.

Definition 2 (Empirical Safe Multi-view Clustering). *When the number of views increases, if the empirical clustering risk of a multi-view model is no higher than that of the model learning from data before view increase and the data of the increased view, this model is said to achieve empirical safe multi-view clustering.*

Moreover, the model should achieve lower empirical risk on unseen data, which motivates us to introduce the following definition.

Definition 3 ($(\epsilon, \delta, \delta_n)$-Expected Safe Multi-view Clustering). *For a given multi-view dataset, when the number of views increases from $p-1$ to p, if for $0 < \delta < 1$, there exists a constant $\epsilon \geq 0$ such that $\mathcal{L} + \epsilon \leq \min\{\mathcal{L}^{\{1,\cdots,p-1\}}, \mathcal{L}^p\} + \delta_n$ holds with at least probability $1-\delta$, where δ_n is a variable related to n which satisfies $\lim_{n\to+\infty} \delta_n = 0$, this model is said to achieve $(\epsilon, \delta, \delta_n)$-expected safe multi-view clustering.*

According to Definition 3, once a model achieves $(\epsilon, \delta, \delta_n)$-expected safe multi-view clustering, its generalization ability is no worse than the model learning from data before view increase and data of the new increased view with high probability. That is, the model is guaranteed to maintain the safe ability even on the data that is unseen in the learning process, which is more applicable in real-world scenarios.

3.3. General Framework of Deep Safe Multi-view Clustering

Generally, a deep MVC model consists of the feature extractor module $\mathcal{F}$ and the cluster assignment module $\mathcal{C}$, where $\mathcal{F}$ and $\mathcal{C}$ can be implemented by the deep neural network. Now we consider the case that the number of views increases from $p-1$ to p. Let $\mathcal{F}_{\{1,\cdots,p-1\}}$ and $\mathcal{F}_{\{1,\cdots,p\}}$ denote the feature extractors of the multi-view model that learns from data before and after view increase. As single-view can be regarded as a special variant of multi-view, this multi-view model can be directly trained from the data of the new increased view. The feature extractor of this single-view variant is denoted as $\mathcal{F}_p$. To simplify the notations, the collection of $\mathcal{F}_p, \mathcal{F}_{\{1,\cdots,p-1\}}, \mathcal{F}_{\{1,\cdots,p\}}$ is denoted as $\{\mathcal{F}\}_p$. To achieve multi-view safeness, we introduce a safe module $\mathcal{S}$ to obtain a combination of the outputs from features extractors

$$\begin{aligned}\mathcal{S}(\{\mathbf{x}^v\}_{v=1}^p; \{\mathcal{F}\}_p) =& \lambda_1 \mathcal{F}_p(\mathbf{x}^p) + \lambda_2 \mathcal{F}_{\{1,\cdots,p-1\}}(\{\mathbf{x}^v\}_{v=1}^{p-1}) \\ &+ \lambda_3 \mathcal{F}_{\{1,\cdots,p\}}(\{\mathbf{x}^v\}_{v=1}^p),\end{aligned} \tag{1}$$

where $\lambda_1, \lambda_2, \lambda_3 \in [0, 1]$ are learnable parameters that represent the safe coefficients assigned by safe module. Then the objective of the proposed deep safe multi-view clustering framework can be formulated as

$$\min_{\boldsymbol{\lambda}\in\Lambda} \left\{ \min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\mathcal{C}(\mathcal{S}(\{\mathbf{x}_i^v\}_{v=1}^p; \{\mathcal{F}\}_p))) \right\}, \tag{2}$$

where $\mathcal{C}$ and $\mathcal{L}$ denote the cluster assignment module and clustering loss, respectively. Θ and Λ include all the parameters in $\{\mathcal{F}_p, \mathcal{F}_{\{1,\cdots,p-1\}}, \mathcal{F}_{\{1,\cdots,p\}}, \mathcal{C}\}$ and $\mathcal{S}$, respectively. To achieve *multi-view safeness*, a constraint $\lambda_1 + \lambda_2 + \lambda_3 = 1$ on $\boldsymbol{\lambda}$ is added. Note that under this constraint, when $\lambda_2 = 1$, the proposed framework degenerates into the multi-view model where only the first $p-1$ views are utilized and the p-th view is discarded. This corresponds to the case that

the new increased view $\mathbf{x}^p$ contains more noise than useful complementary information. Thus the proposed framework should discard the new increased view to eliminate its negative impact on the cluster performance. Besides, when $\lambda_1 = 1$, our proposed framework happens to be the single-view variant. Under this situation, the new increased view contains more useful complementary information than noise, thus directly conducting clustering on the new increased view may achieve better performance. Further, we analyze the empirical risk of the proposed framework and obtain the following theorem,

Theorem 1. *Let the empirical clustering risk of the model learning from data of the new increased view be*

$$\hat{\mathcal{L}}_n^p = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathcal{C}(\mathcal{F}_p(\mathbf{x}_i^p))). \tag{3}$$

The empirical clustering loss of the model learning from data before view increase is denoted as

$$\hat{\mathcal{L}}_n^{\{1,\cdots,p-1\}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathcal{C}(\mathcal{F}_{\{1,\cdots,p-1\}}(\{\mathbf{x}_i^v\}_{v=1}^{p-1}))). \tag{4}$$

Let $\hat{\mathcal{L}}_n^*$ be the optimal value of the optimization problem in Eq. (2). We can prove that $\hat{\mathcal{L}}_n^*$ is no higher than the minimum of $\hat{\mathcal{L}}_n^{\{1,\cdots,p-1\}}$ and $\hat{\mathcal{L}}_n^p$, i.e., $\hat{\mathcal{L}}_n^* \leq \min\{\hat{\mathcal{L}}_n^{\{1,\cdots,p-1\}}, \hat{\mathcal{L}}_n^p\}$.

Proofs of theorems in this paper are provided in the appendix due to the space limit. Theorem 1 shows that our framework can achieve empirical safe multi-view clustering, *i.e.*, Definition 2. That is, the empirical risk of the multi-view model trained by the proposed framework is no higher than that of the model learning from data before view increase and data of the increased view.

3.4. Divergence-based Safe Multi-view Clustering

In Section 3.3, we propose a general framework for deep learning clustering method to achieve *multi-view safeness*, which can be extended to any deep learning based clustering methods by replacing $\{\mathcal{F}_p, \mathcal{F}_{\{1,\cdots,p-1\}}, \mathcal{F}_{\{1,\cdots,p\}}, \mathcal{C}\}$ and $\mathcal{L}$ with specific deep neural network and clustering loss. To verify the effectiveness of our framework, we set the clustering loss $\mathcal{L}$ to the widely used divergence-based loss [12, 36, 44]. Besides, the architecture of $\{\mathcal{F}_p, \mathcal{F}_{\{1,\cdots,p-1\}}, \mathcal{F}_{\{1,\cdots,p\}}, \mathcal{C}\}$ are set to the version adopted in [36, 44]. In this architecture, the output of safe module (*i.e.*, Eq. (1)) is fed to a fully connected layer to obtain hidden features $\mathbf{h}^{(p)}$. Then the cluster assignments $\mathbf{y}^{(p)}$ are obtained from the hidden features by another fully connected layer and a Softmax layer. In this way, we can obtain an example of the proposed framework named Deep Safe Multi-view Clustering (DSMVC). According to Eq. (2), the objective of the proposed DSMVC can be formulated as

$$\min_{\boldsymbol{\lambda}\in\Lambda}\left\{\min_{\boldsymbol{\theta}\in\Theta}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta})\right\}, \tag{5}$$

with

$$\begin{aligned}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta}) =& \frac{1}{\binom{K}{2}}\sum_{l=1}^{K-1}\sum_{s>l}\frac{\mathbf{Y}_{:,l}^{(p)\top}\mathbf{K}^{(p)}\mathbf{Y}_{:,s}^{(p)}}{\sqrt{\mathbf{Y}_{:,l}^{(p)\top}\mathbf{K}^{(p)}\mathbf{Y}_{:,l}^{(p)}\mathbf{Y}_{:,s}^{(p)\top}\mathbf{K}^{(p)}\mathbf{Y}_{:,s}^{(p)}}} \\ &+ \frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j>i}\mathbf{Y}_{i,:}^{(p)}\mathbf{Y}_{j,:}^{(p)\top} \\ &+ \frac{1}{\binom{K}{2}}\sum_{l=1}^{K-1}\sum_{s>l}\frac{\mathbf{D}_{:,l}^{(p)\top}\mathbf{K}^{(p)}\mathbf{D}_{:,s}^{(p)}}{\sqrt{\mathbf{D}_{:,l}^{(p)\top}\mathbf{K}^{(p)}\mathbf{D}_{:,l}^{(p)}\mathbf{D}_{:,s}^{(p)\top}\mathbf{K}^{(p)}\mathbf{D}_{:,s}^{(p)}}}.\end{aligned} \tag{6}$$

$\mathbf{Y}^{(p)} \in \mathbb{R}^{n\times K}$ represents the partition matrix, $\mathbf{Y}_{i,:}^{(p)} = \mathbf{y}_i^{(p)}$. $\mathbf{K}^{(p)} \in \mathbb{R}^{n\times n}$ is the Gaussian kernel matrix, $\mathbf{K}_{ij}^{(p)} = \exp(\|\mathbf{h}_i^{(p)} - \mathbf{h}_j^{(p)}\|^2/(2\sigma^2))$. $\mathbf{D}^{(p)} \in \mathbb{R}^{n\times K}$ represents the similarity matrix between the cluster assignments and the standard simplex $\mathbf{e}_j \in \mathbb{R}^K$, *i.e.*, $\mathbf{D}_{ij}^{(p)} = \exp(\|\mathbf{y}_i^{(p)} - \mathbf{e}_j\|^2)$. The first term in Eq. (6) aims to make the cluster assignments belong to one cluster more compact and that belong to different clusters more separable. The second and the last term are optimized to make the predictions to be orthogonal and close to the standard simplex. The overall learning procedure is summarized as follows. First, the features from data before and after view increase are obtained from $\mathcal{F}_{\{1,\cdots,p-1\}}$ and $\mathcal{F}_{\{1,\cdots,p\}}$. The features from the new increased view are obtained from $\mathcal{F}_p$. Then the hidden features are computed by Eq. (1), *i.e.*, $\mathbf{h}^p = \mathcal{S}(\{\mathbf{x}^v\}_{v=1}^p; \{\mathcal{F}\}_p)$. After that, the objective $\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta})$ is calculated by Eq. (6). Second, for a specific value $\boldsymbol{\lambda}$, the optimal solution of the inner subproblem is obtained by $\boldsymbol{\theta}^*(\boldsymbol{\lambda}) = \operatorname{argmin}_{\boldsymbol{\theta}\in\Theta}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta})$ via gradient descent. Third, the optimal solution of the outer subproblem is obtained by $\boldsymbol{\lambda}^* = \operatorname{argmin}_{\boldsymbol{\lambda}\in\Lambda}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\theta}^*(\boldsymbol{\lambda}))$ via gradient descent. The above process is repeated until convergence. Concretely, for a group of given safe coefficients $\boldsymbol{\lambda}$, the optimization can be regarded as a vanilla deep multi-view clustering process where the model is trained to extract complementary information from multi-view data. Then the safe coefficients are optimized in order to make the model automatically select proper features from the feature extractors. Thus, the whole optimization problem can make the model extract complementary information and discard the meaningless noise by automatically selecting features.

3.5. Theoretical Analysis

Recent work in [28] has made significant breakthrough in deriving sharper generalization bound for kernel and approximate k-means. In [18], a unify theoretical framework for analyzing the generalization of clustering is proposed.

In this part, we study the generalization bound of the proposed DSMVC method. Let $S : \mathcal{X}^2 \mapsto \mathbb{R}_+$ be a similarity function which maps a given pair instances into specific similarity. $H = [H_{1,2}, \cdots, H_{K-1,K}, H_1, \cdots, H_K]$ denotes a collection of partition functions $H_{l,s} : \mathcal{X}^2 \mapsto \mathbb{R}_+$ for $l = 1, \cdots, K-1, s = l+1, \cdots, K$ and $H_l : \mathcal{X}^2 \mapsto \mathbb{R}_+$ for $l = 1, \cdots, K$, which partitions the given instance pairs into disjoint clusters. Then the criterion of the divergence-based clustering framework can be formulated as

$$\begin{aligned}\hat{\mathcal{L}}_n(S,H) =& \frac{1}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s=l+1}^{K} \sum_{i,j=1}^{n} S(\mathbf{x}_i, \mathbf{x}_j) H_{l,s}(\mathbf{x}_i, \mathbf{x}_j) \\ &+ \frac{1}{2\binom{n}{2}} \sum_{l=1}^{K} \sum_{i,j,i\neq j}^{n} H_l(\mathbf{x}_i, \mathbf{x}_j).\end{aligned} \tag{7}$$

Assume that S, $H_{l,s}$ for $l = 1, \cdots, K-1, s = l+1, \cdots, K$ and H_l for $l = 1, \cdots, K$ are symmetry, namely, $S(\mathbf{x}, \mathbf{x}') = S(\mathbf{x}', \mathbf{x})$, $H_{l,s}(\mathbf{x}, \mathbf{x}')$ and $H_l(\mathbf{x}, \mathbf{x}')$ hold for all instance pairs $(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2$. To simplify the notation, $S(\mathbf{x}_i, \mathbf{x}_j)H_{l,s}(\mathbf{x}_i, \mathbf{x}_j)$ and $H_l(\mathbf{x}_i, \mathbf{x}_j)$ are denoted as $g_{S,H_{l,s}}(\mathbf{x}, \mathbf{x}')$ and $g_{H_l}(\mathbf{x}, \mathbf{x}')$, respectively. Let $g_{S,H}$ denote the vector-valued function, *i.e.*, $g_{S,H} := (g_{S,H_{1,2}}, \cdots, g_{S,H_{K-1,K}}, g_{H_1}, \cdots, g_{H_K})$. Then the expectation of Eq. (7) is defined as

$$\begin{aligned}\mathcal{L}(g_{S,H}) =& \mathbb{E}\left[\frac{1}{\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s=l+1}^{K} g_{S,H_{l,s}}(\mathbf{x}, \mathbf{x}')\right] \\ &+ \mathbb{E}\left[\sum_{l=1}^{K} g_{H_l}(\mathbf{x}, \mathbf{x}')\right],\end{aligned} \tag{8}$$

where $\{\mathbf{x}, \mathbf{x}'\}$ is an instance pair sampled from μ. Let $\mathcal{G}$ denote the family of $g_{\kappa,h}$, *i.e.*,

$$\mathcal{G} := \{g_{S,H} | g_{S,H}(\mathbf{x}, \mathbf{x}'), \forall\, \mathbf{x}, \mathbf{x}' \in \mathcal{X}\}. \tag{9}$$

To derive the generation bound, the following assumption is introduced.

Assumption 1. *Assume that the hypothesis functions $g_{S,H_{l,s}}(\cdot,\cdot) \in [0, M]$ for all $l = 1, \cdots, K-1, s > l$ and $g_{H_l}(\cdot,\cdot) \in [0, M]$ for all $l \in [K]$ where $M > 0$ is a constant.*

Remark 1. *In the appendix, we show that the loss function in Eq. (6) is an example of the clustering framework in Eq. (7). Also, it is verified that hypothesis functions corresponding to this loss satisfy Assumption 1.*

It is worth noting that under this clustering framework, the criterion is calculated on a pair of instances, which leads to an independent degree of order $\mathcal{O}(n^2)$. Therefore, directly analyzing the generalization bound via Rademacher complexity is infeasible. To addressed this issue, motivated by [3, 4, 18], the non i.i.d summation form is transformed into a i.i.d summation form by utilizing the permutations in U-process. In this way, the generalization of the proposed deep safe multi-view clustering framework can be derived. Let $\mathcal{L}^p$ and $\mathcal{L}^{\{1,\cdots,p-1\}}$ be the expectation of Eq. (3) and Eq. (4), respectively. The expectation of Eq. (6) is denoted as $\mathcal{L}$. Based on Theorem 1 and Assumption 1, we obtain the following theorem.

Theorem 2. *For any $0 < \delta < 1$, the following inequality holds with at least probability $1 - \delta$:*

$$\mathcal{L} + \epsilon \leq \min\{\mathcal{L}^{\{1,\cdots,p-1\}}, \mathcal{L}^p\} + \frac{c_1}{\sqrt{n}} + c_2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}, \tag{10}$$

where c_1 and c_2 are constants dependent on K and M. ϵ is formulated as $\epsilon := \min\{\hat{\mathcal{L}}_n^p, \hat{\mathcal{L}}_n^{\{1,\cdots,p-1\}}\} - \hat{\mathcal{L}}_n^*$.

We define δ_n as $\delta_n = c_1/\sqrt{n} + c_2/\sqrt{\log(2/\delta)/2n}$. One can see that $\lim_{n\to+\infty} \delta_n = 0$ holds. According to Theorem 1, $\epsilon \geq 0$ holds. Also, ϵ is a constant on a given dataset $\mathcal{D}$. Thus, Theorem 2 shows that the proposed DSMVC can achieve the $(\epsilon, \delta, \delta_n)$-expected safe multi-view clustering, namely, Definition 3. That is, there exists a constant $\epsilon > 0$ such that the expected clustering risk of the proposed DSMVC is at least ϵ lower than that of the model trained from data before view increase and data of the increased view with high probability $1 - \delta$. Thus, the proposed DSMVC is theoretically guaranteed to achieve safe multi-view clustering in terms of both empirical and expected clustering risks. That is to say, on both the training samples and the new unseen samples, our DSMVC is guaranteed to achieve safeness, which could be the best result to achieve *multi-view safeness* under the case that ground-truth labels are not available.

4. Experiments

4.1. Experimental Setup

Datasets. The experiments are conducted on several benchmark multi-view datasets. **Digit** [5] consists of 2,000 instances and each data point is represented by six features, including profile correlations, Fourier coefficients of the character shapes, Karhunen-Love coefficients, morphological features, pixel averages in 2×3 windows, and Zernike moments. **Caltech** [7] is consist of five features from RGB image, including WM, CENTRIST, LBP, GIST, and HOG. We select 1,400 instances from 7 classes and construct a multi-view dataset. **VOC** (*PASCAL VOC 2007*) [6] contains 9,963 image-text pairs from 20 different categories. Following [36, 44], 5,649 instances are selected to construct a two-view dataset, where the first and the second view is 512 Gist features and 399 word frequency count of the instance respectively. **RGB-D** (*SentencesNYUv2*) [14] is an indoor scenes image-text dataset where the image is described by

Dataset	Caltech-2V			Caltech-3V			Caltech-4V			Caltech-5V		
Metric	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
SC [33]	0.567	0.441	0.604	0.625	0.525	0.661	0.692	0.596	0.754	0.772	0.738	0.814
BMVC [43]	0.596	0.445	0.612	0.514	0.462	0.560	0.634	0.537	0.671	0.743	0.676	0.766
RMVC [35]	0.563	0.391	0.574	0.654	0.538	0.665	0.708	0.616	0.746	0.695	0.594	0.731
MVC-LFA [38]	0.462	0.348	0.496	0.551	0.423	0.578	0.609	0.522	0.636	0.741	0.601	0.747
COMIC [31]	0.188	0.147	0.241	0.155	0.134	0.231	0.451	0.573	0.811	0.156	0.111	0.211
EAMC [44]	0.419	0.256	0.427	0.389	0.214	0.398	0.296	0.165	0.310	0.318	0.173	0.342
CoMVC [36]	0.466	0.426	0.527	0.541	0.504	0.584	0.568	0.569	0.646	0.700	0.687	0.729
COMPLETER [22]	0.505	0.509	0.563	0.436	0.440	0.565	0.510	0.514	0.535	0.547	0.550	0.572
OPLFMVC [26]	0.503	0.368	0.520	0.558	0.401	0.567	0.784	0.691	0.806	0.841	0.712	0.841
localized SimpleMKKM [27]	0.567	0.391	0.594	0.664	0.541	0.689	0.739	0.625	0.746	0.700	0.589	0.743
DSMVC (single)	0.564	0.440	0.595	0.598	0.544	0.628	0.656	0.589	0.656	0.871	0.774	0.871
DSMVC (vanilla)	0.533	0.392	0.533	0.622	0.555	0.660	0.767	0.724	0.784	0.841	0.741	0.841
DSMVC	**0.603**	**0.526**	**0.619**	**0.745**	**0.674**	**0.745**	**0.834**	**0.766**	**0.834**	**0.919**	**0.847**	**0.919**

Table 1. Clustering performance comparison on Caltech dataset with increase views. "XV" denotes the number of views.

Dataset	RGB-D			VOC		
Metric	ACC	NMI	Purity	ACC	NMI	Purity
SC [33]	0.312	0.286	0.320	0.372	0.387	0.382
BMVC [43]	0.212	0.082	0.349	0.576	0.535	0.668
RMVC [35]	0.320	0.293	0.348	0.254	0.192	0.294
MVC-LFA [38]	0.415	0.329	0.516	0.503	0.451	0.576
COMIC [31]	0.264	0.131	0.313	0.164	0.435	0.644
EAMC [44]	0.323	0.207	0.311	0.615	0.628	0.615
SiMVC [36]	0.396	0.356	0.344	0.551	0.615	0.740
CoMVC [36]	0.413	0.405	0.413	**0.613**	0.641	0.735
COMPLETER [22]	0.200	0.219	0.421	0.471	0.478	0.574
OPLFMVC [26]	0.416	0.314	0.529	0.580	0.517	0.637
localized SimpleMKKM [27]	0.355	0.273	0.479	0.380	0.303	0.477
DSMVC	**0.431**	**0.416**	**0.602**	0.601	**0.683**	**0.768**

Table 2. Clustering performance comparison on RGB-D and VOC datasets.

the text. The version provided in [36, 44] is adopted in the experiments, which provides visual features from a ResNet-50 network pretrained on the ImageNet dataset and textual features from a doc2vec model pretrained on the Wikipedia dataset. **Multi-MNIST** is a multi-view version of the popular MNIST dataset [16], whose two views are the raw image and its augmented version with a highlighted edge [36, 44].

Baseline Models. We compare DSMVC with several state-of-the art multi-view clustering methods, including Spectral Clustering (SC) [33], BMVC [43], RMVC [35], MVC-LFA [38], COMIC [31], EAMC [44], CoMVC [36], COMPLETER [22], OPLFMVC [26], and localized SimpleMKKM [27]. For spectral clustering, the results on the concatenation of all views are reported as it is a single-view clustering method. To verify the effectiveness of the proposed DSMVC, we report the results of its two versions, including a single-view version (denoted as DSMVC (single)) and a vanilla version (denoted as DSMVC (vanilla)). DSMVC (single) is a single-view model trained from data of the new increased view which corresponds to the case that $\lambda_2 = \lambda_3 = 0$ in Eq. (1). DSMVC (vanilla) is a multi-view model without the proposed safe module, which can be treated as the SiMVC method proposed in [36].

Evaluation Metrics. The clustering performance is evaluated by three metrics: clustering accuracy (ACC), normalized mutual information (NMI), and purity. For all these metrics, a higher value means better performance. Running results with the lowest clustering loss value among 20 independent runs are reported.

Implementation Details. The proposed DSMVC is implemented with PyTorch [30]. The number of training epochs is 120. Following [36, 44], mini-batch gradient descent and Adam optimizer are adopted, and the kernel width σ in Eq. (6) is set to 15% of the median pairwise distance between the semantic features within each mini-batch. Though our theoretical analysis considers the case where the kernel is constructed from overall datasets, we empirically demonstrate that it has little impact due to the well-designed learning schema. Please refer to the appendix for detailed settings of each module.

4.2. Experimental Results

Clustering Performance Comparison. The ACC, NMI, and purity comparison is presented in Table 1, Table 2, Table 3, and Table 4. The clustering results of DSMVC with increasing views are shown in Figure 4. We can obtain the following observations: i) The proposed method outperforms both the single-view variant and the vanilla version (*i.e.*, SiMVC [36]), which is consistent with our analysis in Section 3.3 and Section 3.5. Note that the clustering performance of the vanilla version degrades on Digit dataset when the number of views increases from 4 to 5, which verifies our claim that more views are not always guaranteed to promote the clustering performance. Similarly, the vanilla version performs worse than its single-view variant on Caltech-2V dataset. This result also demonstrates the fact that MVC methods are not guaranteed to perform better than the single-view method. Thus, it is necessary to con-

Dataset	Digit-2V			Digit-3V			Digit-4V			Digit-5V			Digit-6V		
Metric	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
SC [33]	0.647	0.628	0.647	0.643	0.624	0.643	0.628	0.618	0.628	0.647	0.626	0.647	0.663	0.644	0.663
BMVC [43]	0.648	0.624	0.691	0.797	0.814	0.843	0.848	0.827	0.848	0.812	0.835	0.859	0.814	0.859	0.862
RMVC [35]	0.894	0.820	0.894	0.905	0.826	0.905	0.912	0.831	0.912	0.919	0.843	0.919	0.966	0.923	0.966
MVC-LFA [38]	0.797	0.680	0.797	0.658	0.650	0.693	0.782	0.702	0.782	0.787	0.723	0.787	0.646	0.672	0.663
COMIC [31]	0.685	0.757	0.997	0.766	0.844	0.822	0.590	0.705	**0.998**	0.754	0.808	0.912	0.718	0.772	**0.997**
EAMC [44]	0.652	0.610	0.683	0.375	0.289	0.389	0.342	0.258	0.351	0.323	0.226	0.336	0.373	0.240	0.379
CoMVC [36]	0.726	0.737	0.751	0.704	0.749	0.749	0.760	0.791	0.808	0.761	0.765	0.768	0.730	0.799	0.767
COMPLETER [22]	0.651	0.655	0.619	0.761	0.763	0.729	0.622	0.626	0.580	0.652	0.656	0.627	0.792	0.794	0.797
OPLFMVC [26]	0.810	0.690	0.810	0.842	0.724	0.842	0.861	0.754	0.861	0.870	0.766	0.870	0.913	0.829	0.906
localized SimpleMKKM [27]	0.883	0.815	0.885	0.879	0.814	0.886	0.890	0.824	0.896	0.914	0.841	0.915	0.956	0.907	0.956
DSMVC (single)	0.593	0.540	0.596	0.807	0.767	0.807	0.653	0.615	0.654	0.669	0.672	0.686	0.619	0.636	0.623
DSMVC (vanilla)	0.861	0.791	0.861	0.878	0.857	0.878	0.894	0.846	0.894	0.863	0.837	0.863	0.969	0.938	0.969
DSMVC	**0.912**	**0.867**	**0.912**	**0.927**	**0.879**	**0.927**	**0.953**	**0.911**	0.953	**0.960**	**0.914**	**0.960**	**0.978**	**0.950**	0.978

Table 3. Clustering performance comparison on Digit dataset with increase views. "XV" denotes the number of views.

sider both the single-view and multi-view aspects to achieve *multi-view safeness*. Besides, the proposed DSMVC outperforms the rest baseline models including traditional and deep learning based methods in terms of ACC and NMI. This may be attributed to the reason that DSMVC can eliminate the effects of the noise hidden in data by automatically selecting features from single view and multiple views. ii) When the number of views is equal to 2, the proposed DSMVC consists of a multi-view model and two single-view models. As shown in Table 2, on the dataset with two views (*i.e.*, VOC and RGB-D datasets), the proposed DSMVC still outperforms both its vanilla version (*i.e.*, SiMVC [36]) and most baseline methods, which further demonstrates the effectiveness of the proposed safe multi-view learning schema by solving the proposed optimization problem. iii) Our DSMVC surpasses other deep learning based clustering methods on Multi-MNIST dataset, which demonstrates that the proposed safe multi-view mechanism is applicable in large-scale scenarios.

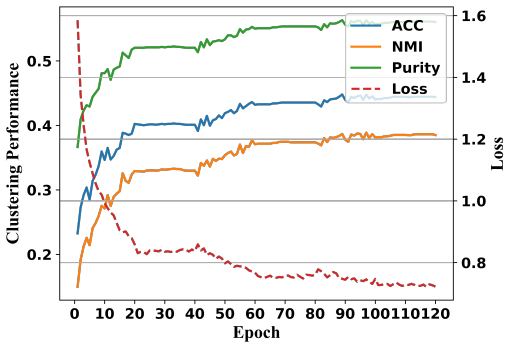


Figure 2. Clustering results of the proposed DSMVC with the increase of iterations on RGB-D.

Visualization and Convergence Analysis. In Figure 2, we plot the clustering loss value and the corresponding clustering performance of DSMVC with iterations to verify its convergence. As observed, the loss value decreases rapidly in the first several epochs. The ACC, NMI, and purity at each iteration are also reported in Figure 2. It can be discovered that the clustering performance of the proposed DSMVC keeps increasing with iterations, and then remains stable. Besides, the hidden features learned by the proposed DSMVC with increasing iterations are visualized by t-SNE [37] in Figure 5. It can be observed that the pro-

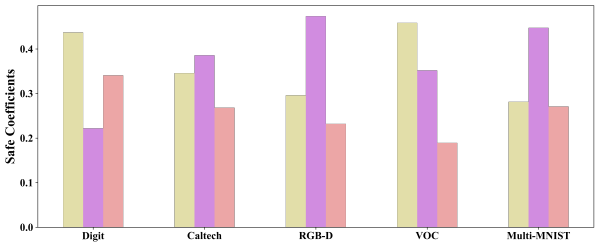


Figure 3. Safe coefficients of DSMVC on all datasets.

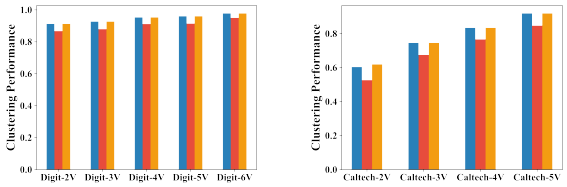
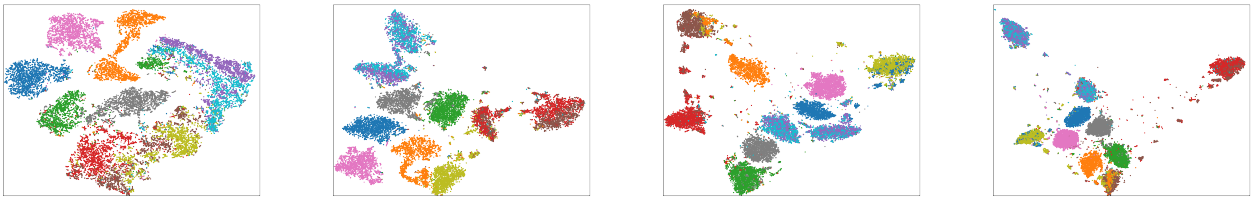


Figure 4. Clustering performance of DSMVC with increasing views on Digit and Caltech.

Dataset	Multi-MNIST		
Metric	ACC	NMI	Purity
EAMC [44]	0.668	0.628	0.651
SiMVC [36]	0.868	0.862	0.870
CoMVC [36]	0.869	0.860	0.869
DSMVC	**0.882**	**0.874**	**0.882**

Table 4. Clustering performance comparison on Multi-MNIST.



(a) Epoch 0 (b) Epoch 40 (c) Epoch 80 (d) Epoch 120

Figure 5. t-SNE visualization of the hidden features on Multi-MNIST dataset with increasing training epochs.

Components			Metric		
$\mathcal{F}_p$	$\mathcal{F}_{\{1,\cdots,p-1\}}$	$\mathcal{F}_{\{1,\cdots,p\}}$	ACC	NMI	Purity
✓			0.871	0.774	0.871
	✓		0.767	0.724	0.784
		✓	0.841	0.741	0.841
✓	✓		0.889	0.803	0.889
	✓	✓	0.731	0.675	0.742
✓		✓	0.901	0.830	0.901
✓	✓	✓	0.919	0.847	0.919

Table 5. Ablation study on Caltech dataset. "✓" in the table represents DSMVC with the corresponding component.

posed DSMVC can learn a more compact and separated cluster structure during training. All these observations demonstrate the effectiveness of the proposed safe multi-view learning mechanism.

Safe Coefficients Analysis. We then investigate the safe coefficients (*i.e.*, λ_1, λ_2, and λ_3) learned by the proposed DSMVC, whose values are presented in Figure 3. It can be seen that the proposed DSMVC can learn a group of non-sparse safe coefficients on all datasets. As discussed in Section 3.3, a group of sparse safe coefficients means that the model degenerates to the single-view variant (*i.e.*, DSMVC (single)) or the vanilla variant (*i.e.*, DSMVC (vanilla)). Note that the learned safe coefficients are the optimal solution of the outer subproblem according to Eq. (2). Based on Theorem 2, a group of non-sparse safe coefficients corresponds to the case that the expected clustering risk of DSMVC is lower than that of the single view variant and the vanilla variant with high probability. Indeed, the observations on all datasets show that DSMVC does perform better, which demonstrates that DSMVC is consistent in theory and experiments.

Ablation Study. In this section, we design an ablation study to demonstrate the superiority of the proposed safe multi-view mechanism. As mentioned in Section 3.3, the proposed deep safe multi-view clustering framework contains three feature extractors, including a single-view extractor which receives data of the new increased view and two multi-view feature extractors which receive data before and after view increase respectively. Thus, there are seven combinations between these three feature extractors and the cluster assignment module. We experimentally evaluate these combinations on Caltech dataset with all the views, *i.e.*, Caltech-5V dataset. As shown in Table 5, the proposed DSMVC outperforms all other combinations, which indicates that each component in the proposed framework is contributed to the clustering performance.

5. Conclusion

In this paper, we address the safeness of multi-view clustering under the case where views dynamically increase. By the proposed optimization problem, our framework can extract complementary information and discard the meaningless noise. In theory, the empirical clustering risk of the proposed DSMVC is no higher than learning from data before the view increased and data of the new increased view. And the expected clustering risk of the proposed DSMVC is no higher than that with high probability. We believe that our work will bring more insights in improving the robustness of multi-view clustering on real-world scenarios. The main limitation of the proposed DSMVC is lacking efficient skill of the proposed optimization problem. Future work may focus on designing more effective optimization approach, and extend our framework to incomplete multi-view data.

Acknowledgements

The authors sincerely appreciate the anonymous reviewers and area chairs for their helpful and invaluable comments. This work is supported in part by the National Natural Science Foundation of China (No.62076234, No.61703396, No. 62106257), Beijing Outstanding Young Scientist Program NO.BJJWZYJH012019100020098, Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" initiative, Renmin University of China, China Unicom Innovation Ecological Cooperation Plan, Public Computing Cloud of Renmin University of China, Beijing Natural Science Foundation (No. 4222029).

References

[1] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–594, 2015. 2

[2] Man-Sheng Chen, Ling Huang, Chang-Dong Wang, and Dong Huang. Multi-view clustering in latent embedding space. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 3513–3520, 2020. 2

[3] Stéphan Clemençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and scoring using empirical risk minimization. In *International Conference on Computational Learning Theory (COLT 2005)*, pages 1–15. Springer, 2005. 5

[4] Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008. 5

[5] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. 5

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 5

[7] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004. 5

[8] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3897–3906, 2020. 2

[9] Pengxin Guo, Feiyang Ye, and Yu Zhang. Safe multi-task learning. *arXiv preprint arXiv:2111.10601*, 2021. 2

[10] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021. 2

[11] Zhenyu Huang, Joey Tianyi Zhou, Xi Peng, Changqing Zhang, Hongyuan Zhu, and Jiancheng Lv. Multi-view spectral clustering network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2563–2569, 2019. 2

[12] Michael Kampffmeyer, Sigurd Løkse, Filippo M Bianchi, Lorenzo Livi, Arnt-Børre Salberg, and Robert Jenssen. Deep divergence-based approach to clustering. *Neural Networks*, 113:91–101, 2019. 4

[13] Zhao Kang, Wangtao Zhou, Zhitong Zhao, Junming Shao, Meng Han, and Zenglin Xu. Large-scale multi-view subspace clustering in linear time. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 4412–4419, 2020. 2

[14] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3565, 2014. 5

[15] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems*, pages 1413–1421, 2011. 2

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6

[17] Ruihuang Li, Changqing Zhang, Huazhu Fu, Xi Peng, Tianyi Zhou, and Qinghua Hu. Reciprocal multi-layer subspace learning for multi-view clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8172–8180, 2019. 2

[18] Shaojie Li and Yong Liu. Sharper generalization bounds for clustering. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6392–6402, 2021. 4, 5

[19] Yu-Feng Li, Han-Wen Zha, and Zhi-Hua Zhou. Learning safe prediction for semi-supervised regression. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 2217–2223, 2017. 2

[20] Yu-Feng Li, Lan-Zhe Guo, and Zhi-Hua Zhou. Towards safe weakly supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):334–346, 2019. 2

[21] Yu-Feng Li and Zhi-Hua Zhou. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):175–188, 2014. 2

[22] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11174–11183, 2021. 6, 7

[23] Jiyuan Liu, Xinwang Liu, Siwei Wang, Sihang Zhou, and Yuexiang Yang. Hierarchical multiple kernel clustering. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 8671–8679, 2021. 2

[24] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Li Liu, Siqi Wang, Weixuan Liang, and Jiangyong Shi. One-pass multi-view clustering for large-scale data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12344–12353, 2021. 2

[25] Xinwang Liu, Yong Dou, Jianping Yin, Lei Wang, and En Zhu. Multiple kernel k-means clustering with matrix-induced regularization. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 1888–1894, 2016. 2

[26] Xinwang Liu, Li Liu, Qing Liao, Siwei Wang, Yi Zhang, Wenxuan Tu, Chang Tang, Jiyuan Liu, and En Zhu. One pass late fusion multi-view clustering. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6850–6859, 2021. 2, 6, 7

[27] Xinwang Liu, Sihang Zhou, Li Liu, Chang Tang, Siwei Wang, Jiyuan Liu, and Yi Zhang. Localized simple multiple kernel k-means. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9293–9301, 2021. 2, 6, 7

[28] Yong Liu. Refined learning bounds for kernel and approximate k-means. In *Advances in Neural Information Processing Systems*, 2021. 4

[29] Feiping Nie, Jing Li, and Xuelong Li. Self-weighted multiview clustering with multiple graphs. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2564–2570, 2017. 2

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019. 6

[31] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5092–5101, 2019. 2, 6, 7

[32] Yazhou Ren, Shudong Huang, Peng Zhao, Minghao Han, and Zenglin Xu. Self-paced and auto-weighted multi-view clustering. *Neurocomputing*, 383:248–256, 2020. 2

[33] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 6, 7

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

[35] Hong Tao, Chenping Hou, Xinwang Liu, Tongliang Liu, Dongyun Yi, and Jubo Zhu. Reliable multi-view clustering. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 4123–4130, 2018. 1, 2, 6, 7

[36] Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1255–1265, 2021. 1, 2, 4, 5, 6, 7

[37] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 7

[38] Siwei Wang, Xinwang Liu, En Zhu, Chang Tang, Jiyuan Liu, Jingtao Hu, Jingyuan Xia, and Jianping Yin. Multi-view clustering via late fusion alignment maximization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3778–3784, 2019. 2, 6, 7

[39] Jie Xu, Yazhou Ren, Guofeng Li, Lili Pan, Ce Zhu, and Zenglin Xu. Deep embedded multi-view clustering with collaborative training. *Information Sciences*, 573:279–290, 2021. 2

[40] Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9234–9243, October 2021. 2

[41] Changqing Zhang, Huazhu Fu, Qinghua Hu, Xiaochun Cao, Yuan Xie, Dacheng Tao, and Dong Xu. Generalized latent multi-view subspace clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):86–99, 2018. 2

[42] Chen Zhang, Siwei Wang, Jiyuan Liu, Sihang Zhou, Pei Zhang, Xinwang Liu, En Zhu, and Changwang Zhang. Multi-view clustering via deep matrix factorization and partition alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 4156–4164, 2021. 2

[43] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1774–1782, 2018. 2, 6, 7

[44] 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 14619–14628, 2020. 1, 2, 4, 5, 6, 7

[45] Sihang Zhou, Xinwang Liu, Jiyuan Liu, Xifeng Guo, Yawei Zhao, En Zhu, Yongping Zhai, Jianping Yin, and Wen Gao. Multi-view spectral clustering with optimal neighborhood laplacian matrix. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 6965–6972, 2020. 2

[46] Xiaofeng Zhu, Shichao Zhang, Wei He, Rongyao Hu, Cong Lei, and Pengfei Zhu. One-step multi-view spectral clustering. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):2022–2034, 2019. 2