Deep Safe Incomplete Multi-view Clustering: Theorem and Algorithm

Huayi Tang [1 2] **Yong Liu** [1 2]

Abstract

Incomplete multi-view clustering is a significant but challenging task. Although jointly imputing incomplete samples and conducting clustering has been shown to achieve promising performance, learning from both complete and incomplete data may be worse than learning only from complete data, particularly when imputed views are semantic inconsistent with missing views. To address this issue, we propose a novel framework to reduce the clustering performance degradation risk from semantic inconsistent imputed views. Concretely, by the proposed bi-level optimization framework, missing views are dynamically imputed from the learned semantic neighbors, and imputed samples are automatically selected for training. In theory, the empirical risk of the model is no higher than learning only from complete data, and the model is never worse than learning only from complete data in terms of expected risk with high probability. Comprehensive experiments demonstrate that the proposed method achieves superior performance and efficient safe incomplete multi-view clustering.

1. Introduction

Multi-view data, containing modalities from multiple domains, exists widely in real-world application scenarios. For example, multiple types of information is provided by sensors attached to the autonomous vehicle, which are treated as multiple views. Due to the expensive cost of collecting a large amount of data with manual annotations, numerous studies in multi-view clustering (Nie et al., 2016; Zhang et al., 2017; Peng et al., 2019; Liu et al., 2021c; Xu et al., 2021; Pan & Kang, 2021; Xu et al., 2022b) are developed and demonstrate that mining the complementary information of multiple views yields promising performance.

The aforementioned studies are based on the assumption that instances contain the same number of views, which may not be satisfied in real-world applications. Taking autonomous driving as an example, some types of information from sensors is missing due to hardware faults or interference signals, leading to the incompleteness of multi-view data. Recent years have witnessed the development of incomplete multi-view clustering (IMVC) approaches (Li et al., 2014; Shao et al., 2015; Zhao et al., 2016; Hu & Chen, 2018; Guo & Ye, 2019; Wen et al., 2020; Lin et al., 2021; Xu et al., 2022a), and most of them fall into imputation approaches that jointly fill incomplete instances and conduct clustering. However, learning from complete and filled samples is sometimes worse than learning only from complete data. Intuitively, the model attempts to recover the missing views without ground-truth information, which may affect the quality of imputed views. Cluster-oriented imputed samples that are semantic consistent with the missing samples boost clustering performance, yet the semantic inconsistency between imputed views and other views disturbs the intrinsic common semantics of multiple views, leading to the difficulty of learning consistent cluster assignments and degenerated clustering performance. Therefore, incomplete multi-view clustering should focus on the following two challenges at the same time, namely (i) how to achieve semantic consistency between imputed views and missing views? and (ii) how to reduce the risk of clustering performance degradation caused by semantic inconsistency between imputed views and missing views? Although existing studies have explored to learn imputations with high semantic consistency, efforts to simultaneously address these two challenges, particularly reducing the cluster performance degradation risk caused by semantic inconsistent imputed views, are still limited.

To this end, we propose a novel IMVC framework named Deep Safe Incomplete Multi-View Clustering (DSIMVC) to achieve safe incomplete multi-view clustering, namely, learning from both complete and incomplete data is no worse than learning only from complete data. Towards this goal, a weighting function is introduced to automatically assign weights to the incomplete samples. On the one hand, the weighting function is optimized to minimize the empirical clustering risk of the learner on complete data.

[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. Correspondence to: Yong Liu <liuyonggsai@ruc.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

On the other hand, the model is trained from complete data and weighted incomplete data, which reduces the negative effects from low-quality imputed views, especially with semantic inconsistency. These two learning processes are cast as a unified bi-level optimization framework. Besides, DSIMVC dynamically mines the k-nearest neighbors based on learned semantic features, from which missing views are imputed. By this means, the learned representation features guide neighbor search and missing views imputation, which further promotes the model to learn better representation features. In theory, on complete data, the empirical clustering risk of DSIMVC is no higher than learning only from complete data. Also, the expected clustering risk of DSIMVC is no higher than learning only from complete data with high probability. Experimental results on public datasets demonstrate the superiority and effectiveness of the proposed learning schema.

2. Related Work

In this section, we briefly introduce the recent development of the topics related to our work, including IMVC and safeness studies in machine learning.

Incomplete multi-view clustering. Existing incomplete multi-view clustering methods can be divided into traditional methods (Liu et al., 2020; Zhang et al., 2021; Li et al., 2022) and deep learning based methods (Xu et al., 2019; Wang et al., 2021a; Zhang et al., 2022; Yang et al., 2022). In (Li et al., 2014), common latent subspace is mined via non-negative matrix factorization technique. In (Wen et al., 2019), latent features from multiple views are aligned, and the common local structure is exploited via a consensus graph. Collaboratively imputing incomplete kernel matrices and conducting clustering are first introduced in (Liu et al., 2020). By a well-designed self-paced learning based framework, the work (Wen et al., 2020) reduces the negative influence of the marginal samples. Motivated from the information theory, a unified framework is proposed in (Lin et al., 2021) to jointly learn consistent representation and recover the missing view by maximizing the mutual information while minimizing the conditional entropy of multiple views. The work (Liu et al., 2021b) proposes to impute the incomplete base matrix from multiple views with a learned consensus matrix regularized by prior knowledge. A one-stage late fusion method is introduced in (Zhang et al., 2021) that incorporates the imputation of missing views and clustering. The work (Wang et al., 2021b) proposes to generate the missing views via graph neural network based on the inter-instance relationships that are transferred from other views. Learnable latent representation is introduced in (Zhang et al., 2022) to mine the common semantics from multiple views. The authors in (Yang et al., 2022) establish a unified framework to address view-aligned and sample-missing problems. The differences between previous approaches and our work are as follows. First, the neighbors are mined from the learned representation features and updated dynamically, in contrast to (Wang et al., 2021b) where the relationships are based on raw input and remain unchanged. Second, the proposed framework is theoretically guaranteed to achieve no degraded clustering performance.

Safeness studies in machine learning. Safeness studies in machine learning aim to reduce the risk of performance degradation. For semi-supervised and weakly supervised learning, safeness means that the performance of a learner does not degrade by using unlabeled data. In (Li & Zhou, 2014), multiple low-density separators are utilized to approximate the ground-truth decision boundary. The work (Li et al., 2017) proposes a geometric projection based framework to learn predictions from multiple semi-supervised regressors. Further, in (Li et al., 2019), a unified learning schema is proposed where the ground-truth label assignment is approximated by a convex linear combination of base weakly supervised learners. In (Guo et al., 2020), a deep learning based framework that tackles the performance degradation caused by class mismatch via bi-level optimization is presented. For unsupervised learning, the work (Tao et al., 2018) establishes a min-max optimization based framework to guarantee that multi-view methods are no worse than a given single-view method. Recent work (Tang & Liu, 2022) achieves multi-view safeness where the number of views dynamically increases. The differences between previous studies and our work are summarized as follows. First, the works (Li & Zhou, 2014; Li et al., 2017; 2019) focus on semi-supervised or weakly supervised learning where partial ground-truth labels are available, while our work focus on clustering where all ground-truth labels are not available. Second, different from the work (Tao et al., 2018) that relies on sample completeness assumption, the proposed framework is feasible for data with missing views. This work is inspired by (Ren et al., 2018; Shu et al., 2019; Guo et al., 2020). It should be pointed out that the goal of (Guo et al., 2020) is eliminating performance degradation caused by class mismatch in semi-supervised learning, yet our work aims to provide not degenerated performance for model learning from both complete and incomplete data in IMVC. Another main difference lies in the theoretical results, *i.e.*, we demonstrate that learning with complete and imputed samples is not worse than learning only with complete data in terms of expected risk with high probability under the proposed framework.

3. Methodology

In this section, we first give the notations and definitions used in this paper. Then we detail the proposed deep safe incomplete multi-view clustering framework. After that, we

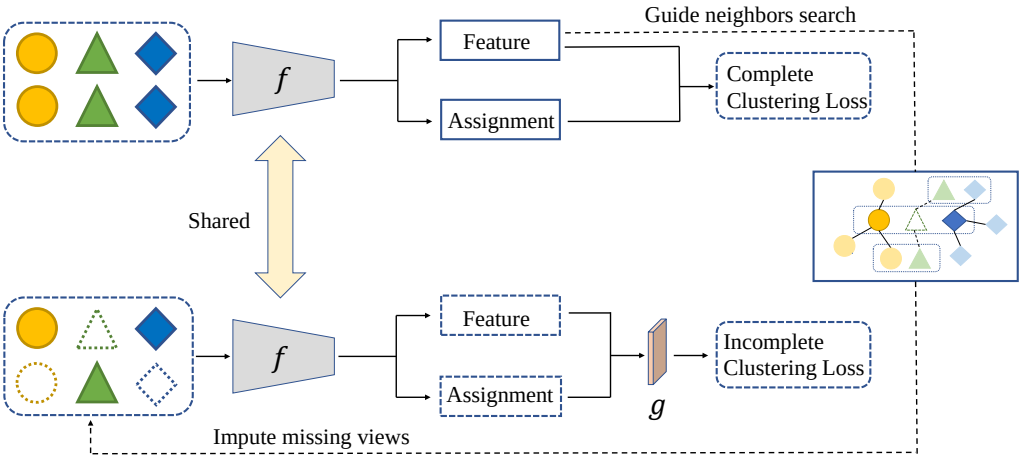


Figure 1. Overview of the DSIMVC framework. Each row represents a sample and different shapes indicate multiple views. Edges of available and missing views are indicated by solid and dashed lines, respectively.

present the theoretical analysis of our framework, including convergence analysis and its mechanism to achieve safe incomplete multi-view clustering.

3.1. Notation and Definition

The incomplete multi-view dataset with n instances sampled i.i.d. from input space $\mathcal{X}$ is denoted as $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$, where each instance contains m views and the p-th view of sample $\mathbf{x}$ is denoted as $\mathbf{x}^p$. The existence of views are described by an indicator matrix $\mathbf{M} \in \{0,1\}^{n\times m}$, *i.e.*, $\mathbf{M}_{ip} = 1$ denotes the p-th view of the i-th sample is available, otherwise $\mathbf{M}_{ip} = 0$. $\binom{n}{2}$ denotes the combination number. The number of complete and incomplete samples are denoted as n_c and n_e, respectively. Let K be the number of categories that is known in advance. To reduce the risk of clustering performance degradation from incomplete data, the clustering performance of the model learning from both complete and incomplete data should be no worse than learning only from complete data. However, due to all ground-truth labels are not available, it is hard to measure the clustering performance of the model. According to the empirical risk minimization, the model should minimize the empirical clustering risk on complete data that contains complete common semantic information. With this observation in mind, we present the following definition.

Definition 3.1 (Safe Incomplete Multi-view Clustering)**.** For a given multi-view dataset, if the empirical clustering risk on the complete data of the model learning from both complete and incomplete data is no higher than learning only from complete data, this model is said to achieve empirical safe incomplete multi-view clustering. Further, a model is defined to achieve expected safe incomplete multi-view clustering if its expected clustering risk is no higher than learning only from complete data with high probability.

Thus, our goal is to build a new incomplete multi-view clustering framework with theoretical guarantee to achieve the defined safe incomplete multi-view clustering.

3.2. Deep Safe Incomplete Multi-view Clustering

Let $f : \mathcal{X} \mapsto \mathbb{R}^D \times \mathbb{R}^K$ denotes the function that maps input samples into semantic features and cluster assignment probability. In this work, f is implemented by a deep neural network with parameters w. Then, for a given sample $\mathbf{x}_i^p$, its semantic features and cluster assignment probability are denoted as $f_{\mathcal{Z}}(\mathbf{x}_i^p; w) \in \mathbb{R}^D$ and $f_{\mathcal{Q}}(\mathbf{x}_i^p; w) \in \mathbb{R}^K$, respectively. Due to the superior ability of deep neural networks in learning representation, the geometric relationships of feature vectors reflect the semantic relationships of samples to some extent (Van Gansbeke et al., 2020). That is, samples may belong to the same category if their feature vectors are close to each other in feature space, which motivates us to recover the missing views from neighbors inferred by features. Note that with the increase of iterations, features with more semantic information are learned and more reliable neighbors can be mined. Thus, neighbors are dynamically updated to better describe the intrinsic relationships of samples. Inspired by (Zhong et al., 2021), the semantic features of the available view $\mathbf{x}_i^p$ (*i.e.*, $\mathbf{M}_{ip} = 1$) in the t-th iteration is updated in a moving-average manner:

$$\mathbf{z}_i^{p,t} = \frac{(1-\gamma)\mathbf{z}_i^{p,t-1} + \gamma f_{\mathcal{Z}}(\mathbf{x}_i^p; w)}{\|(1-\gamma)\mathbf{z}_i^{p,t-1} + \gamma f_{\mathcal{Z}}(\mathbf{x}_i^{p,t}; w)\|_2},$$

where γ is the trade-off coefficient. According to the fact that multiple views of a sample share common semantic information, the neighbors of other views serve as complementary information to find the neighbors of the current view. Thus, semantic neighbors of the sample $\mathbf{x}_i^p$ in the t-th

iteration are defined as

$$\mathcal{N}_i^{p,t} := \bigcup_{q=1,q\neq p}^{m} \left\{ \mathbf{x}_j^p \middle| j \in \Psi_i^{q,t} \right\}, \tag{1}$$

with

$$\Psi_i^{q,t} := \left\{ j \middle| \mathbf{z}_j^{q,t} \in \mathcal{N}^k(\mathbf{z}_i^{q,t}), \mathbf{M}_{ip} = \mathbf{M}_{jq} = \mathbf{M}_{jp} = 1 \right\},$$

where $\Psi_i^{q,t}$ denotes the neighbors' indices of $\mathbf{x}_i^p$ that inferred from $\mathbf{x}_i^q$ in the t-th iteration, and $\mathcal{N}^k(\mathbf{z}^{q,t})$ represents the k-nearest neighbors of the semantic features $\mathbf{z}^{q,t}$. By this way, the semantic information learned from features among all views is utilized to find more reliable neighbors, from which the missing are imputed. In this work, we simply impute the missing views with the average of semantic neighbors, *i.e.*,

$$\hat{\mathbf{x}}_i^{p,t} = \frac{1}{|\mathcal{N}_i^{p,t}|} \sum_{\mathbf{x}^p \in \mathcal{N}_i^{p,t}} \mathbf{x}^p. \tag{2}$$

After that, imputed views and other views are constructed as a subset $\mathcal{D}^e := \{\tilde{\mathbf{x}}_i^t\}_{i=1}^{n_e}$, where the p-th view of $\tilde{\mathbf{x}}^t$ is denoted as $\tilde{\mathbf{x}}^{p,t}$. For each incomplete sample, missing views are imputed while other views are retained, *i.e.*, $\tilde{\mathbf{x}}_i^{p,t} = \hat{\mathbf{x}}_i^{p,t}$ if $\mathbf{M}_{ip} = 0$ otherwise $\tilde{\mathbf{x}}_i^{p,t} = \mathbf{x}_i^p$. It is worth noticing that $\mathcal{D}^e$ is updated in each iteration to improve the semantic consistency between imputed views and missing views. Since imputed views are inferred from learned neighbors, the semantic consistency between imputed views and missing views depends on the reliability of learned neighbors. We resort to the following objective based on spectral contrastive loss (HaoChen et al., 2021) to achieve the alignment among views and mine high-quality features:

$$\begin{aligned}\mathcal{L}_F(f(\mathcal{D}^c; w)) = &\sum_{p=1}^{m}\sum_{q=p+1}^{m}\Bigg[-\frac{2}{n_c}\sum_{i=1}^{n_c} f_{\mathcal{Z}}(\mathbf{x}_i^p; w)^\top f_{\mathcal{Z}}(\mathbf{x}_i^q; w) \\ &+ \frac{1}{2\binom{n_c}{2}}\sum_{i=1}^{n_c}\sum_{j\neq i}\left(f_{\mathcal{Z}}(\mathbf{x}_i^p; w)^\top f_{\mathcal{Z}}(\mathbf{x}_j^q; w)\right)^2\Bigg].\end{aligned}$$

Besides, since multiple views contain common semantics, the cluster assignment probabilities among views should be consistent. Thus, the following objective is utilized to align the predictions from multiple views:

$$\begin{aligned}\mathcal{L}_C(f(\mathcal{D}^c; w)) = -\frac{1}{K}\sum_{p=1}^{m}\sum_{q=p+1}^{m}\sum_{j=1}^{K}\Bigg[&\log\frac{e^{\mathcal{Q}_j^{p\top}\mathcal{Q}_j^q}}{\sum_{s\neq j} e^{\mathcal{Q}_j^{p\top}\mathcal{Q}_s^p}} \\ &+ \log\frac{e^{\mathcal{Q}_j^{p\top}\mathcal{Q}_j^q}}{\sum_{s\neq j} e^{\mathcal{Q}_j^{q\top}\mathcal{Q}_s^q}}\Bigg],\end{aligned}$$

where $\mathcal{Q}^p = [f_{\mathcal{Q}}(\mathbf{x}_i^p; w)^\top; \cdots; f_{\mathcal{Q}}(\mathbf{x}_{n_c}^p; w)^\top] \in \mathbb{R}^{n_c\times K}$ and $\mathcal{Q}_j^p$ is the j-th column of $\mathcal{Q}^p$. Following (Huang et al., 2020; Van Gansbeke et al., 2020; Zhong et al., 2021), a regularization term is jointly optimized to prevent the trivial solution, which is formulated as

$$\mathcal{L}_R(f(\mathcal{D}^c; w)) = \sum_{p=1}^{m}\sum_{j=1}^{K}\bar{\mathcal{Q}}_j^p \log \bar{\mathcal{Q}}_j^p,$$

where $\bar{\mathcal{Q}}_j^p = \frac{1}{n_c}\sum_{i=1}^{n_c}\mathcal{Q}_{ij}^p$. The clustering loss on complete data is

$$\begin{aligned}\mathcal{L}(f(\mathcal{D}^c; w) =&\mathcal{L}_F(f(\mathcal{D}; w)) + \mathcal{L}_C(f(\mathcal{D}; w)) \\ &+ \mathcal{L}_R(f(\mathcal{D}; w)).\end{aligned}$$

The clustering loss $\mathcal{L}(f(\mathcal{D}^e; w)$ on filled incomplete data is defined accordingly by substituting $\mathbf{x}$ for $\tilde{\mathbf{x}}$. However, the model may still face the risk of performance degradation when inferred neighbors are semantic inconsistent with missing views. Therefore, DSIMVC dynamically selects incomplete samples for training via a weighting function $g : \mathcal{X} \mapsto \mathbb{R}_+$ with parameters ϕ. To achieve the empirical safe incomplete multi-view clustering in Definition 3.1, we propose the following optimization problem:

$$\begin{aligned}&\min_{\phi,w} \mathcal{L}(f(\mathcal{D}^c; w)) \quad s.t. \quad w \in \mathcal{S}(\phi) \\ &\mathcal{S}(\phi) = \operatorname*{argmin}_{w} \mathcal{L}(f(\mathcal{D}^c; w)) + \mathcal{L}(f(\mathcal{D}^e; w), g(\mathcal{D}^e; \phi)),\end{aligned} \tag{3}$$

where

$$\begin{aligned}&\mathcal{L}(f(\mathcal{D}^e; w), g(\mathcal{D}^e; \phi)) \\ :=&\sum_{i=1}^{n_e} g(\tilde{\mathbf{x}}_i; w)\Bigg[-\frac{2}{n_e}\sum_{p=1}^{m}\sum_{q=p+1}^{m} f_{\mathcal{Z}}(\tilde{\mathbf{x}}_i^p; w)^\top f_{\mathcal{Z}}(\tilde{\mathbf{x}}_i^q; w) \\ &+ \frac{1}{2\binom{n_e}{2}}\sum_{j\neq i}\left(f_{\mathcal{Z}}(\tilde{\mathbf{x}}_i^p; w)^\top f_{\mathcal{Z}}(\tilde{\mathbf{x}}_j^q; w)\right)^2 \\ &+ \mathcal{L}_C(f(\mathcal{D}^e; w)) + \mathcal{L}_R(f(\mathcal{D}^e; w))\Bigg].\end{aligned}$$

$\mathcal{D}^c := \{\mathbf{x}_i\}_{i=1}^{n_c}$ is the subset of $\mathcal{D}$ with complete samples, and $\mathcal{L} : \mathbb{R}^D \times \mathbb{R}^K \mapsto \mathbb{R}$ represents the clustering loss. Eq. (3) is a bi-level optimization problem that contains two levels of optimization tasks (Sinha et al., 2017), *i.e.*, the lower-level and the upper-level optimization problems. Concretely, the lower-level problem is a traditional multi-view clustering problem that aims to find the best multi-view model f learning from both complete data and incomplete data with weights given by g. In the upper-level optimization problem, the weighting function g is optimized such that the model returned by the lower-level optimization task achieves the lowest empirical risk on complete data, by which cluster-beneficial filled incomplete instances are selected. Solving Eq. (3) is challenging since the global optima is arduous to obtain (Liu et al., 2021a; Bao et al., 2021). To this end, we assume that the lower-level singleton condition holds (Franceschi et al., 2018; Ren et al., 2018), which implies that the optimal solution of the upper-level

optimization task is unique and reduces the difficulty in solving the bi-level optimization problem. With this assumption, Eq. (3) is reformulated as

$$\begin{aligned} &\min_{\phi,w} \mathcal{L}(f(\mathcal{D}^c; w^*(\phi))) \\ w^*(\phi) = &\operatorname*{argmin}_{w} \mathcal{L}(f(\mathcal{D}^c; w)) + \mathcal{L}(f(\mathcal{D}^e; w), g(\mathcal{D}^e; \phi)). \end{aligned} \tag{4}$$

Eq. (4) can be solved by gradient based optimization approach (Shu et al., 2019; Guo et al., 2020). Specifically, in each training iteration, the optimal solution of the lower-level optimization problem (*i.e.*, $w^*(\phi)$) is approximated by the one-step iterative value $\hat{w}^{(t)}(\phi)$ towards the gradient descent direction, namely,

$$\begin{aligned} \hat{w}^{(t)}(\phi) =& w^{(t)} - \eta_w \nabla_w \mathcal{L}(f(\mathcal{D}^c; w^{(t)})) \\ &- \eta_w \nabla_w \mathcal{L}(f(\mathcal{D}^e; w^{(t)}), g(\mathcal{D}^e; \phi)). \end{aligned}$$

After that, the approximate solution of the lower-level optimization problem is transmitted into the upper-level optimization problem to guide the update of the weighting network parameters, *i.e.*,

$$\phi^{(t+1)} = \phi^{(t)} - \eta_\phi \nabla_\phi \mathcal{L}(f(\mathcal{D}^c; \hat{w}^{(t)}(\phi^{(t)}))). \tag{5}$$

In this step, the weighting network automatically changes the weights of incomplete samples to minimize the empirical risk of the learner f on complete samples. Based on the new weighted incomplete samples, parameters of the learner f are updated by

$$\begin{aligned} w^{(t+1)} =& w^{(t)} - \eta_w \nabla_w \mathcal{L}(f(\mathcal{D}^c; w^{(t)})) \\ &- \nabla_w \mathcal{L}(f(\mathcal{D}^e; w^{(t)}), g(\mathcal{D}^e; \phi^{(t+1)})). \end{aligned} \tag{6}$$

By this mechanism, those high-quality imputed samples (*i.e.*, semantic consistent samples) are selected and those low-quality imputed samples (*i.e.*, semantic inconsistent samples) are discarded, which reduces the clustering performance degradation risk caused by semantic inconsistency. The overall learning process is summarized in Algorithm 1.

3.3. Theoretical Analysis

We first analyze the convergence of the proposed framework. To simplify the notations, the objective of the upper-level optimization problem in Eq. (4) is denoted as $\mathcal{L}(w(\phi))$. According to the aforementioned optimization procedure, we have the following theorem.

Theorem 3.2. *Suppose that $g(\cdot;\phi)$ and loss function $\mathcal{L}(\cdot,\cdot;w)$ are twice differential with bound gradients and Hessians. Suppose that the learning rate η_w satisfies $\eta_w = \min\{1, \frac{k}{T}\}$ for some $k > 0$ such that $\frac{k}{T} < 1$ and $\eta_\phi = \min\{\frac{1}{L}, \frac{C}{\sqrt{T}}\}$ for some $C > 0$, such that $\frac{\sqrt{T}}{C} \geq L$. Then the proposed bi-level optimization problem can achieve $\min_{0\leq t\leq T} \mathbb{E}[\|\nabla \mathcal{L}(\hat{w}^{(t)}(\phi^{(t)}))\|_2^2] \leq \epsilon$ in $\mathcal{O}(1/\epsilon^2)$ steps.*

Algorithm 1 Deep Safe Incomplete Multi-view Clustering

Input: Incomplete multi-view data $\{\mathbf{x}_1^p, \ldots, \mathbf{x}_n^p\}_{p=1}^m$, number of cluster K, learning rate η_w and η_ϕ, max iterations T.
Output: Cluster assignments $\hat{\mathbf{y}}$.
Initialize the parameters of f, g, semantic features $\{\mathbf{z}^{p,0}\}_{p=1}^m$, and semantic neighbors $\{\mathcal{N}^{p,0}\}_{p=1}^m$.
for $t = 0$ **to** $T-1$ **do**
  Impute the incomplete views by Eq. (2).
  Sample a random mini-batch complete data from $\mathcal{D}^c$ and incomplete data from $\mathcal{D}^e$.
  Compute the lower-level objective by Eq. (4).
  Update ϕ by Eq. (5).
  Update w by Eq. (6).
  Update semantic neighbors according to Eq. (1).
end for
Compute the overall cluster assignment probability matrix by $\mathcal{Q} = \frac{1}{m}\sum_{p=1}^m \mathcal{Q}^p$.
Compute cluster assignments by $\hat{\mathbf{y}}_i = \operatorname{argmax}_j \mathcal{Q}_{ij}$.

Proofs of theorems in this paper are provided in the appendix due to space limit. Theorem 3.2 demonstrates that the optimization algorithm theoretically converges to the (local) optima. Next, to see that the proposed framework can achieve empirical safe multi-view clustering, we analyze the empirical clustering risk of DSIMVC and obtain the following theorem.

Theorem 3.3. *Let $\widehat{\mathcal{L}}(f(\mathcal{D}; w))$ be the empirical clustering risk on complete data $\mathcal{D}^c$. The parameters of the multi-view model learning only from complete data and the optimal solution of Eq. (4) are denoted as $w^* = \operatorname{argmin}_{w\in\mathcal{W}} \widehat{\mathcal{L}}(f(\mathcal{D}^c; w))$ and $\hat{\phi}$ respectively. We can prove that $\widehat{\mathcal{L}}(f(\mathcal{D}^c; w^*(\hat{\phi}))) \leq \widehat{\mathcal{L}}(f(\mathcal{D}^c; w^*))$.*

Theorem 3.3 reveals that the multi-view learner is theoretically guaranteed to achieve empirical safe incomplete multi-view clustering in Definition 3.1 under the proposed bi-level optimization framework, *i.e.*, the empirical clustering risk of DSIMVC is no higher than that of the model learning only from complete data. We further analyze the ability of DSIMVC to achieve safe incomplete multi-view clustering on unseen data. Let $\widehat{\mathcal{L}}(f(\mathcal{D}^c; w^*(\hat{\phi}))$ be the empirical clustering risk of DSIMVC and $\widehat{\mathcal{L}}(f(\mathcal{D}^c; w^*))$ be the empirical clustering risk of the model learning only from complete data. The expectation of $\widehat{\mathcal{L}}(f(\mathcal{D}^c; w^*(\hat{\phi}))$ and $\widehat{\mathcal{L}}(f(\mathcal{D}^c; w^*))$ are denoted as $\mathcal{L}(f(w^*(\hat{\phi})))$ and $\mathcal{L}(f(w^*))$, respectively. The family of f is defined as $\mathcal{F}$. Recent works (Liu, 2021; Li & Liu, 2021) establish pioneering theoretical analysis for sharper generalization bound of clustering approaches. Inspired by these studies, we obtain the following theorem by analyzing the generalization bound of the proposed DSIMVC method.

Theorem 3.4. *Suppose that* $\|f_{\mathcal{Z}}(\mathbf{x}^p)\|_\infty \leq E$ *hold for all* $\{\mathbf{x}^p\}_{p=1}^m \in \mathcal{X}$*, where* $E > 0$ *is a constant. For any* $0 < \delta < 1$*, with at least probability* $1-\delta$ *for any* $f \in \mathcal{F}$*, the following inequality holds*

$$\mathcal{L}(f(w^*(\hat{\phi}))) + \varepsilon \leq \mathcal{L}(f(w^*)) + \frac{c_1}{\sqrt{n_c}} + c_2\sqrt{\frac{\log 12/\delta}{n_c}},$$

where c_1 *and* c_2 *are constants dependent on* D, E, K, m*.* ε *is formulated as* $\varepsilon := \widehat{\mathcal{L}}(f(\mathcal{D}; w^*)) - \widehat{\mathcal{L}}(f(\mathcal{D}; w^*(\hat{\phi}))$.

According to Theorem 3.2, we have $\varepsilon \geq 0$. Theorem 3.4 shows that the proposed framework can achieve expected safe incomplete multi-view clustering in Definition 3.1. That is, with high probability $1-\delta$, the expected clustering risk on the complete part of DSIMVC is no higher than learning only from complete data. To summarize, the proposed framework is theoretically guaranteed to achieve safe incomplete multi-view clustering in terms of both empirical and generalization clustering risk, which may be the best guarantee for safe incomplete clustering where all ground-truth labels are not available.

4. Experiments

4.1. Experimental Setup

Datasets. The experiments are conducted on several widely used benchmark multi-view datasets. **BDGP** (Cai et al., 2012) is a drosophila embryos image dataset with 2,500 samples of 5 objects, where each sample is described by 1750-D visual feature and 79-D textual feature. **MNIST-USPS** (Peng et al., 2019) contains 5,000 samples of 10 categories, where the first view and the second view are sampled from the popular MNIST (LeCun et al., 1998) and USPS handwritten digit datasets, respectively. **Columbia Consumer Video (CCV)** (Jiang et al., 2011) is composed of 6,773 samples from 20 categories. 5,000-D STIP features, 5,000-D SIFT, and 4,000-D MFCC features extracted from YouTube videos are treated as three views. **Multi-Fashion** is a two-view dataset constructing from Fashion-MNIST (Xiao et al., 2017) that consists of 5,000 samples. Following the same manner to construct MNIST-USPS, the original image and the randomly sampled image from the same category are regarded as two views.

Baseline methods. We compare the proposed framework with the following baselines: best single view clustering (BSV), PVC (Li et al., 2014), UEAF (Wen et al., 2019), CDIMC-net (Wen et al., 2020), MKKM-IK (Liu et al., 2020) COMPLETER (Lin et al., 2021), EE-R-IMVC (Liu et al., 2021b), and OS-IF-IMVC (Zhang et al., 2021). Following (Zhao et al., 2016), missing views are first imputed by the average value of available views and then the best results obtained by k-means among all views are reported in BSV.

Evaluation metrics. The clustering performance is evaluated by three metrics, including clustering accuracy (ACC), normalized mutual information (NMI), and purity. For all these metrics, a higher value means better performance. The experiment on each dataset is repeated 10 times independently and the average values and the standard deviations are reported.

Implementation details. For each dataset, we generate incomplete samples by randomly removing views under the condition that at least one view remained in the sample. The ratio of incomplete sample sizes to overall sample sizes is denoted as p, which ranges from 0.1 to 0.7 with 0.2 as interval. The implementation is based on PyTorch (Paszke et al., 2019) platform. Please refer to the appendix for the experiment details and results of the purity comparison.

4.2. Experimental Results

Clustering performance comparison. The ACC and NMI comparison is presented in Table 1. From this table, we obtain the following observations: (i) Overall, the other IMVC methods perform better than BSV, which indicates that the imputed views inferred by other IMVC methods contain more semantic information than the average vectors and thus alleviate the clustering risk degradation risk caused by semantic inconsistency. (ii) The proposed DSIMVC significantly outperforms the other methods on all datasets, especially with high dimensional input features and more incomplete views. For example, on CCV with a missing ratio of 0.7, DSIMVC exceeds the second best one by about 7.5% and 9.2% in terms of ACC and NMI, respectively. This result demonstrates the superiority of jointly mining semantic imputed views and reducing the clustering performance risk. (iii) When the missing ratio increased from 0.5 to 0.7 on BDGP, compared with the second best one whose ACC decrease by 10.2% in terms of ACC, the ACC of our method decrease by only 3.2%, which demonstrates the effectiveness of the proposed learning schema to achieve safe incomplete multi-view clustering by automatically weighting imputed samples. These observations demonstrate the superiority of DSIMVC against other methods, which is due to the proposed bi-level optimization based learning schema reducing negative effects of semantic inconsistent filled views.

Convergence analysis and visualization. In Figure 2, we plot the objective value and the values of evaluation metrics with iterations to verify its convergence. One can observe that the objective value decreases rapidly and then continuously decrease until convergence. Also, the values of ACC, NMI, and purity firstly increase with iterations and then keep fluctuation in a narrow range. These results demonstrate the convergence of DSIMVC, which is consistent with the theoretical analysis in Theorem 3.2. Afterward, to verify the effectiveness of DSIMVC in mining semantic neighbors,

Table 1. Clustering accuracy (ACC) and normalized mutual information (NMI) comparison (mean±std) of different methods on all benchmark datasets with different missing ratios. Best results are shown in bold.

		ACC				NMI			
Dataset	Method\p	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
BDGP	BSV	59.64±1.43	54.67±1.34	44.49±0.59	35.96±0.32	47.21±1.54	42.53±1.40	32.24±0.63	22.97±0.37
	PVC	55.64±2.48	54.99±0.13	68.33±7.49	59.87±0.26	29.37±2.43	31.53±0.12	46.69±6.97	42.25±0.51
	UEAF	90.66±0.57	89.32±0.00	87.08±0.00	76.88±0.00	74.87±1.33	71.99±0.00	67.15±0.00	53.75±0.00
	CDIMC-net	80.47±0.82	74.67±0.53	67.71±1.05	56.11±4.80	70.08±0.36	67.64±0.78	54.51±1.12	39.70±4.85
	MKKM-IK	65.01±0.03	59.80±0.00	52.56±0.00	43.84±0.00	49.62±0.51	35.22±0.00	24.55±0.00	14.58±0.00
	EE-R-IMVC	65.28±0.00	57.36±0.00	42.48±0.00	34.85±2.48	43.82±0.00	31.79±0.00	21.39±0.00	11.87±1.99
	COMPLETER	40.91±7.04	41.80±4.12	41.54±7.64	39.63±2.78	33.19±4.33	31.15±5.43	32.62±5.58	27.47±3.37
	OS-LF-IMVC	82.78±2.18	74.34±1.16	59.71±3.22	45.34±1.39	60.25±4.69	48.27±2.33	30.56±3.97	18.54±1.31
	DSIMVC	**98.40±0.26**	**96.93±0.45**	**95.29±0.37**	**92.14±0.84**	**94.67±0.91**	**90.34±1.13**	**86.11±0.92**	**79.37±1.56**
MNIST-USPS	BSV	49.15±1.76	42.57±1.70	35.62±1.67	26.67±1.04	45.15±0.69	39.17±0.78	31.73±0.94	23.62±0.41
	PVC	64.57±2.73	63.04±3.69	52.56±1.14	50.24±2.84	58.74±1.66	55.63±1.03	46.35±0.47	44.34±1.33
	UEAF	71.27±0.97	66.08±1.26	61.94±0.00	54.18±0.00	66.75±1.81	58.04±2.14	57.84±0.00	49.77±0.00
	CDIMC-net	52.23±4.52	49.72±1.10	47.97±1.13	31.78±1.68	61.45±2.74	64.40±2.42	56.62±0.87	34.79±0.83
	MKKM-IK	72.25±0.61	64.44±0.00	49.74±1.04	35.70±0.00	61.64±0.18	52.01±0.00	37.67±0.59	24.68±0.00
	EE-R-IMVC	75.07±0.50	58.86±0.00	45.58±0.00	28.02±0.00	64.27±0.17	49.47±0.00	34.15±0.00	16.97±0.00
	COMPLETER	96.87±1.04	96.56±0.82	93.66±5.63	83.80±6.05	93.94±1.29	92.31±1.18	90.51±2.71	81.18±2.94
	OS-LF-IMVC	62.29±1.80	46.58±2.93	32.83±1.45	23.70±0.86	49.14±2.42	33.98±2.11	22.22±0.82	13.96±0.66
	DSIMVC	**98.88±0.09**	**97.89±0.14**	**96.78±0.25**	**93.34±0.64**	**96.91±0.21**	**94.50±0.36**	**91.98±0.55**	**85.64±0.93**
CCV	BSV	18.91±0.37	17.55±0.41	15.74±0.26	14.46±0.27	17.22±0.15	15.61±0.20	13.44±0.15	11.46±0.10
	PVC	16.48±0.40	15.54±0.27	14.75±0.33	14.01±0.24	13.86±0.36	10.12±0.28	9.67±0.27	8.66±0.18
	UEAF	26.38±0.00	24.82±0.00	22.63±0.00	14.92±3.20	23.64±0.00	23.10±0.00	21.34±0.00	10.42±3.66
	CDIMC-net	18.53±1.10	18.20±1.24	17.41±0.56	14.53±0.98	15.88±0.68	14.89±0.72	13.45±1.06	9.28±1.12
	MKKM-IK	19.71±0.38	18.29±0.00	15.46±0.00	14.13±0.00	14.78±0.06	12.61±0.00	10.30±0.00	8.00±0.00
	EE-R-IMVC	25.29±0.04	23.03±0.00	17.87±0.00	14.78±0.00	21.43±0.10	17.53±0.00	12.35±0.00	7.48±0.00
	COMPLETER	21.72±1.30	20.62±0.48	18.38±0.73	17.35±0.69	22.57±0.96	19.59±0.66	17.33±0.80	13.73±0.79
	OS-LF-IMVC	20.47±0.74	17.15±0.63	14.21±0.50	12.37±0.46	15.34±0.56	12.23±0.36	9.50±0.37	7.05±0.46
	DSIMVC	**30.90±1.22**	**29.33±1.24**	**27.07±0.81**	**24.87±0.49**	**29.76±0.71**	**28.18±0.65**	**25.72±0.61**	**22.96±0.56**
Multi-Fashion	BSV	49.81±2.60	42.97±2.01	34.83±1.32	26.59±0.83	48.32±0.99	40.85±0.60	32.46±0.64	23.73±0.41
	PVC	45.69±0.44	40.77±1.50	42.01±2.61	40.55±0.79	44.98±0.33	39.32±1.07	39.78±1.12	39.2±0.71
	UEAF	57.07±0.67	50.88±2.88	48.96±0.88	30.34±0.00	57.15±1.72	48.79±4.78	44.04±4.03	24.13±0.00
	CDIMC-net	51.00±4.89	44.73±2.23	42.10±3.00	37.61±3.68	62.52±1.94	54.67±1.94	44.85±4.19	46.05±1.29
	MKKM-IK	70.08±0.12	59.96±0.00	46.38±0.00	29.84±0.00	61.29±0.13	50.52±0.00	38.25±0.00	20.64±0.00
	EE-R-IMVC	72.83±0.97	63.32±0.00	51.16±0.00	20.24±0.00	65.78±0.36	57.28±0.00	43.50±0.00	14.61±0.00
	COMPLETER	78.63±0.33	71.68±3.70	70.76±5.62	69.33±4.51	82.23±1.18	77.12±0.57	74.76±1.35	70.23±2.73
	OS-LF-IMVC	62.54±1.01	50.10±2.68	37.47±1.38	27.67±1.25	52.36±1.11	38.74±2.05	30.04±1.48	19.98±1.70
	DSIMVC	**89.60±0.89**	**87.47±1.23**	**83.79±1.40**	**75.71±1.69**	**84.47±0.70**	**81.76±1.07**	**77.82±0.73**	**71.53±1.45**

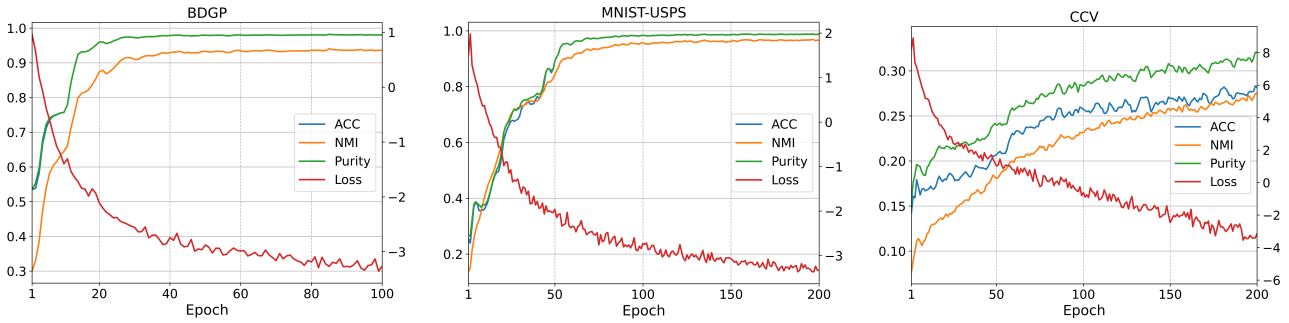
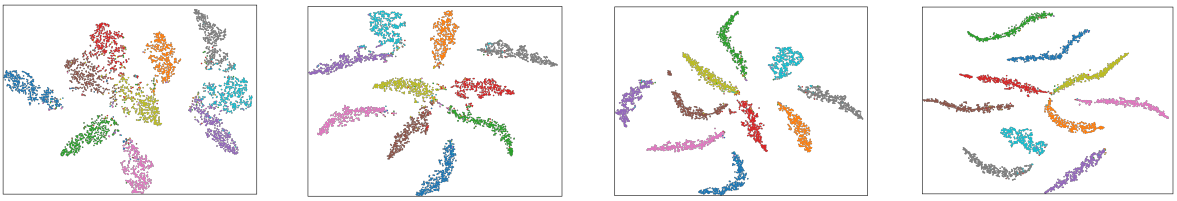


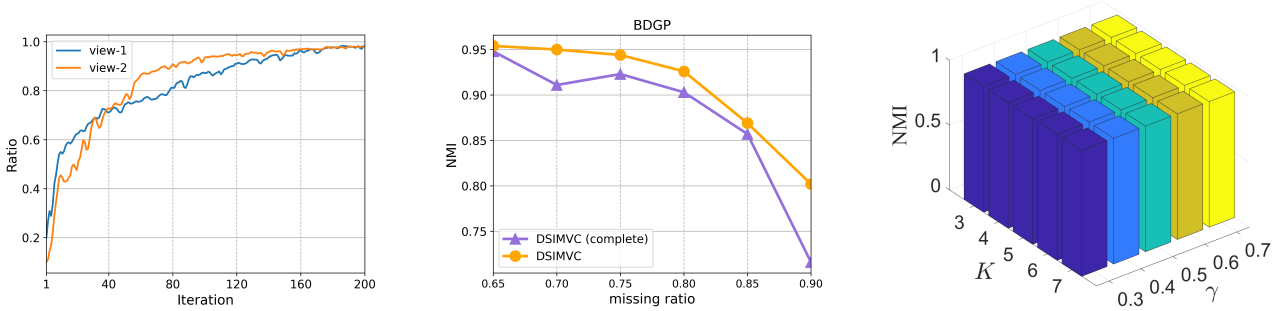
Figure 2. The objective value and clustering performance of DSIMVC with the increase of iterations on BDGP, MNIST-USPS, and CCV.

we further calculated the proportion of semantic consistent samples to available samples in each view of MNIST-USPS, where the semantic consistent sample means its category is the same as the average of inferred neighbors' categories. Figure 4(a) illustrates that, due to the proposed framework dynamically assigning semantic consistent imputations with higher scores and semantic inconsistent imputations with lower scores, the quality of learned representation is improved and thus leads to an increasing number of semantic consistent samples. Besides, the learned features with increasing iterations are visualized by t-SNE (Van der Maaten & Hinton, 2008). As shown in Figure 3, the cluster struc-



(a) Iteration 50 (b) Iteration 100 (c) Iteration 150 (d) Iteration 200

Figure 3. t-SNE visualization of the learned features on MNIST-USPS with increasing training iterations.



(a) (b) (c)

Figure 4. Model analysis. (a) Semantic consistency ratio of the learned neighbors with the increase of iterations; (b) Clustering performance in terms of NMI of DSIMVC and its variant with different missing ratios; (c) Parameters sensitivity analysis.

ture becomes more compact and separated with increasing iterations, which corresponds to a higher semantic consistency ratio. The aforementioned observations verify that the proposed framework alleviates the clustering degradation risk from semantic inconsistent imputed views and thus improves the clustering quality.

Parameter analysis and ablation study. In this part, we first conduct experiments to evaluate the effect of the hyper-parameters on clustering performance, and then evaluate the effectiveness of the proposed framework to achieve safe incomplete multi-view clustering. The hyper-parameters of DSIMVC include the number of neighbors k and the trade-off coefficient γ. Figure 4(c) presents the NMI of DSIMVC by varying k from 3 to 7 and γ from 0.3 to 0.7. As observed, the clustering performance of DSIMVC is insensitive with both k and γ in a wide range, which indicates that our framework is insensitive to the variation of the hyper-parameters. Thus, γ and k are empirically set to 0.5 and 3, respectively. Besides, we evaluate a special variant of DSIMVC that learns only from complete data (denoted as DSIMVC (complete)) on BDGP with different missing ratios, and the clustering performance on complete samples are presented in Figure 4(b). One can find that the clustering performance of DSIMVC is no worse than its variant on complete data, which is consistent with theoretical results. Observations on other datasets are similar apart from some cases where the performance is affected by the approximations adopted in solving the bi-level problem and inferior local optima. These observations demonstrate that the proposed bi-level optimization based framework reduces the clustering performance degradation risk effectively.

5. Conclusion

In this paper, we propose a unified framework with theoretical guarantee to simultaneously mine semantic consistent imputations and reduce the clustering performance degradation risk from semantic inconsistent imputations. By the proposed bi-level optimization framework, missing views are dynamically imputed from semantic neighbors, and the incomplete samples are automatically selected for learning. In theory, the empirical clustering risk on complete data of the model is never higher than learning only from complete data. Also, with high probability, the model is no worse than learning only from complete data in terms of generalization risk. Experimental results on public datasets demonstrate the effectiveness of the proposed framework in achieving safe incomplete multi-view clustering. We hope our work could bring new insights to recover high-quality imputed views and improve the robustness of multi-view learners on incomplete data. Our future work includes developing new approaches to solve the bi-level optimization problem and extending the proposed framework to more challenging scenarios where all samples contain missing views.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive comments to improve the quality of this paper. 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

Bao, F., Wu, G., Li, C., Zhu, J., and Zhang, B. Stability and generalization of bilevel programming in hyperparameter optimization. In *Advances in Neural Information Processing Systems*, pp. 4529–4541, 2021.

Cai, X., Wang, H., Huang, H., and Ding, C. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12):i16–i24, 2012.

Clémençon, S., Lugosi, G., and Vayatis, N. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1568–1577, 2018.

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

Guo, L.-Z., Zhang, Z.-Y., Jiang, Y., Li, Y.-F., and Zhou, Z.-H. Safe deep semi-supervised learning for unseen-class unlabeled data. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3897–3906, 2020.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, pp. 5000–5011, 2021.

Hu, M. and Chen, S. Doubly aligned incomplete multi-view clustering. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2262–2268, 2018.

Huang, J., Gong, S., and Zhu, X. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8849–8858, 2020.

Jiang, Y.-G., Ye, G., Chang, S.-F., Ellis, D., and Loui, A. C. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, pp. 1–8, 2011.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7 (3):535–547, 2019.

Latała, R. and Oleszkiewicz, K. On the best constant in the khinchin-kahane inequality. *Studia Mathematica*, 109(1): 101–104, 1994.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, S. and Liu, Y. Sharper generalization bounds for clustering. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6392–6402, 2021.

Li, S.-Y., Jiang, Y., and Zhou, Z.-H. Partial multi-view clustering. *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 28(1), 2014.

Li, Y.-F. and Zhou, Z.-H. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):175–188, 2014.

Li, Y.-F., Zha, H.-W., and Zhou, Z.-H. Learning safe prediction for semi-supervised regression. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 2217–2223, 2017.

Li, Y.-F., Guo, L.-Z., and Zhou, Z.-H. Towards safe weakly supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):334–346, 2019.

Li, Z., Tang, C., Zheng, X., Liu, X., Zhang, W., and Zhu, E. High-order correlation preserved incomplete multi-view subspace clustering. *IEEE Transactions on Image Processing*, 31:2067–2080, 2022.

Lin, Y., Gou, Y., Liu, Z., Li, B., Lv, J., and Peng, X. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11174–11183, 2021.

Liu, R., Gao, J., Zhang, J., Meng, D., and Lin, Z. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021a.

Liu, X., Zhu, X., Li, M., Wang, L., Zhu, E., Liu, T., Kloft, M., Shen, D., Yin, J., and Gao, W. Multiple kernel k-means with incomplete kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1191–1204, 2020.

Liu, X., Li, M., Tang, C., Xia, J., Xiong, J., Liu, L., Kloft, M., and Zhu, E. Efficient and effective regularized incomplete multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2634–2646, 2021b.

Liu, X., Liu, L., Liao, Q., Wang, S., Zhang, Y., Tu, W., Tang, C., Liu, J., and Zhu, E. One pass late fusion multi-view clustering. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6850–6859, 2021c.

Liu, Y. Refined learning bounds for kernel and approximate k-means. In *Advances in Neural Information Processing Systems*, pp. 6142–6154, 2021.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT press, 2018.

Nie, F., Li, J., and Li, X. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 1881–1887, 2016.

Pan, E. and Kang, Z. Multi-view contrastive graph clustering. In *Advances in Neural Information Processing Systems*, pp. 2148–2159, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8026–8037, 2019.

Peng, X., Huang, Z., Lv, J., Zhu, H., and Zhou, J. T. Comic: Multi-view clustering without parameter selection. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5092–5101, 2019.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4334–4343, 2018.

Shao, W., He, L., and Yu, P. S. Multiple incomplete views clustering via weighted nonnegative matrix factorization with $L_{2,1}$ regularization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 318–334, 2015.

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., and Meng, D. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, 2019.

Sinha, A., Malo, P., and Deb, K. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22(2):276–295, 2017.

Tang, H. and Liu, Y. Deep safe multi-view clustering: Reducing the risk of clustering performance degradation caused by view increase. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 202–211, 2022.

Tao, H., Hou, C., Liu, X., Liu, T., Yi, D., and Zhu, J. Reliable multi-view clustering. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 4123–4130, 2018.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., and Van Gool, L. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pp. 268–285. Springer, 2020.

Wang, Q., Ding, Z., Tao, Z., Gao, Q., and Fu, Y. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30:1771–1783, 2021a.

Wang, Y., Chang, D., Fu, Z., and Zhao, Y. Incomplete multi-view clustering via cross-view relation transfer. *arXiv preprint arXiv:2112.00739*, 2021b.

Wen, J., Zhang, Z., Xu, Y., Zhang, B., Fei, L., and Liu, H. Unified embedding alignment with missing views inferring for incomplete multi-view clustering. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 5393–5400, 2019.

Wen, J., Zhang, Z., Xu, Y., Zhang, B., Fei, L., and Xie, G.-S. CDIMC-net: Cognitive deep incomplete multi-view clustering network. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pp. 3230–3236, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xu, C., Guan, Z., Zhao, W., Wu, H., Niu, Y., and Ling, B. Adversarial incomplete multi-view clustering. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3933–3939, 2019.

Xu, J., Ren, Y., Tang, H., Pu, X., Zhu, X., Zeng, M., and He, L. 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*, pp. 9234–9243, 2021.

Xu, J., Li, C., Ren, Y., Peng, L., Mo, Y., Shi, X., and Zhu, X. Deep incomplete multi-view clustering via mining cluster complementarity. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2022a.

Xu, J., Tang, H., Ren, Y., Peng, L., Zhu, X., and He, L. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16051–16060, 2022b.

Yang, M., Li, Y., Hu, P., Bai, J., Lv, J. C., and Peng, X. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Zhang, C., Hu, Q., Fu, H., Zhu, P., and Cao, X. Latent multi-view subspace clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4279–4287, 2017.

Zhang, C., Cui, Y., Han, Z., Zhou, J. T., Fu, H., and Hu, Q. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2402–2415, 2022.

Zhang, Y., Liu, X., Wang, S., Liu, J., Dai, S., and Zhu, E. One-stage incomplete multi-view clustering via late fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2717–2725, 2021.

Zhao, H., Liu, H., and Fu, Y. Incomplete multi-modal visual data grouping. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 2392–2398, 2016.

Zhong, H., Wu, J., Chen, C., Huang, J., Deng, M., Nie, L., Lin, Z., and Hua, X.-S. Graph contrastive clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9224–9233, 2021.

A. Proofs.

In this appendix, we provide the detailed proofs of the theoretical results.

A.1. Proof of Theorem 3.2

The proof is motivated by (Shu et al., 2019). For concise, the multi-view sample is denoted as z. We first show that the upper-level objective as a function of ϕ is Lipschitz smooth. The update of w is formulated as

$$\begin{aligned} w^{(t+1)} =& w^{(t)} - \frac{\eta_w}{n_c}\sum_{i=1}^{n_c}\left(\nabla_w \ell^{(1)}(z_i;w) + \nabla_w \ell^{(2)}(z_i;w)\right) - \frac{\eta_w}{n_c(n_c-1)}\sum_{i=1}^{n_c}\sum_{j\neq i}\nabla_w \ell(z_i,z_j;w) \\ & - \frac{\eta_w}{n_c}\sum_{i=1}^{n_c} g(\tilde{z}_i;\phi^{(t)})\left(\nabla_w \ell^{(1)}(z_i;w) + \nabla_w \ell^{(2)}(z_i;w)\right) - \frac{\eta_w}{n_c(n_c-1)}\sum_{i=1}^{n_c} g(\tilde{z}_i;\phi^{(t)})\left(\sum_{j\neq i}\nabla_w \ell(z_i,z_j;w)\right), \end{aligned}$$

where

$$\begin{aligned} \nabla_w \ell^{(1)}(z_i;w) :=& -2\sum_{p=1}^{m}\sum_{q=p+1}^{m}\left[\nabla_w f_{\mathcal{Z}}(\mathbf{x}_i^p;w)^\top f_{\mathcal{Z}}(\mathbf{x}_i^q;w) + \nabla_w f_{\mathcal{Z}}(\mathbf{x}_i^q;w)^\top f_{\mathcal{Z}}(\mathbf{x}_i^p;w)\right] \\ \nabla_w \ell(z_i,z_j;w) :=& 2\sum_{p=1}^{m}\sum_{q=p+1}^{m}\left(f_{\mathcal{Z}}(\mathbf{x}_i^p;w)^\top f_{\mathcal{Z}}(\mathbf{x}_j^q;w)\right)\times \\ & \left[\nabla_w f_{\mathcal{Z}}(\mathbf{x}_i^p;w)^\top f_{\mathcal{Z}}(\mathbf{x}_j^q;w) + \nabla_w f_{\mathcal{Z}}(\mathbf{x}_j^q;w)^\top f_{\mathcal{Z}}(\mathbf{x}_i^p;w)\right] \\ \nabla_w \ell^{(2)}(z_i;w) :=& \frac{1}{K}\sum_{p=1}^{m}\sum_{q=p+1}^{m}\sum_{l=1}^{K}\Bigg[-\left(f_{\mathcal{Q}}^l(\mathbf{x}_i^p;w)\nabla_w f_{\mathcal{Q}}^l(\mathbf{x}_i^q;w) + f_{\mathcal{Q}}^l(\mathbf{x}_i^q;w)\nabla_w f_{\mathcal{Q}}^l(\mathbf{x}_i^p;w)\right) \\ & + \frac{1}{\sum_{s\neq j} e^{\mathcal{Q}_j^{p\top}\mathcal{Q}_s^p}}\sum_{k\neq l} e^{\mathcal{Q}_l^{p\top}\mathcal{Q}_k^p}\left(f_{\mathcal{Q}}^l(\mathbf{x}_i^p;w)\nabla_w f_{\mathcal{Q}}^k(\mathbf{x}_i^p;w) + f_{\mathcal{Q}}^k(\mathbf{x}_i^p;w)\nabla_w f_{\mathcal{Q}}^l(\mathbf{x}_i^p;w)\right) \\ & + \frac{1}{\sum_{s\neq j} e^{\mathcal{Q}_j^{q\top}\mathcal{Q}_s^q}}\sum_{k\neq l}\left(f_{\mathcal{Q}}^l(\mathbf{x}_i^q;w)\nabla_w f_{\mathcal{Q}}^k(\mathbf{x}_i^q;w) + f_{\mathcal{Q}}^k(\mathbf{x}_i^q;w)\nabla_w f_{\mathcal{Q}}^l(\mathbf{x}_i^q;w)\right)\Bigg] \\ & + \sum_{p=1}^{m}\sum_{l=1}^{K}\left[\log\bar{\mathcal{Q}}_l^p + 1\right]\left(\nabla_w f_{\mathcal{Q}}^l(\mathbf{x}_i^p)\right), \end{aligned}$$

and

$$\begin{aligned} \nabla_w \ell^{(1)}(\tilde{z}_i;w) :=& -2\sum_{p=1}^{m}\sum_{q=p+1}^{m}\left[\nabla_w f_{\mathcal{Z}}(\tilde{\mathbf{x}}_i^p;w)^\top f_{\mathcal{Z}}(\tilde{\mathbf{x}}_i^q;w) + \nabla_w f_{\mathcal{Z}}(\tilde{\mathbf{x}}_i^q;w)^\top f_{\mathcal{Z}}(\tilde{\mathbf{x}}_i^p;w)\right] \\ \nabla_w \ell(\tilde{z}_i,\tilde{z}_j;w) :=& 2\sum_{p=1}^{m}\sum_{q=p+1}^{m}\left(f_{\mathcal{Z}}(\tilde{\mathbf{x}}_i^p;w)^\top f_{\mathcal{Z}}(\tilde{\mathbf{x}}_j^q;w)\right)\times \\ & \left[\nabla_w f_{\mathcal{Z}}(\tilde{\mathbf{x}}_i^p;w)^\top f_{\mathcal{Z}}(\tilde{\mathbf{x}}_j^q;w) + \nabla_w f_{\mathcal{Z}}(\tilde{\mathbf{x}}_j^q;w)^\top f_{\mathcal{Z}}(\tilde{\mathbf{x}}_i^p;w)\right] \\ \nabla_w \ell^{(2)}(\tilde{z}_i;w) :=& \frac{1}{K}\sum_{p=1}^{m}\sum_{q=p+1}^{m}\sum_{l=1}^{K}\Bigg[-2\left(f_{\mathcal{Q}}^l(\tilde{\mathbf{x}}_i^p;w)\nabla_w f_{\mathcal{Q}}^l(\tilde{\mathbf{x}}_i^q;w) + f_{\mathcal{Q}}^l(\tilde{\mathbf{x}}_i^q;w)\nabla_w f_{\mathcal{Q}}^l(\tilde{\mathbf{x}}_i^p;w)\right) \\ & + \frac{1}{\sum_{s\neq j} e^{\tilde{\mathcal{Q}}_j^{p\top}\tilde{\mathcal{Q}}_s^p}}\sum_{k\neq l} e^{\tilde{\mathcal{Q}}_l^{p\top}\tilde{\mathcal{Q}}_k^p}\left(f_{\mathcal{Q}}^l(\tilde{\mathbf{x}}_i^p;w)\nabla_w f_{\mathcal{Q}}^k(\tilde{\mathbf{x}}_i^p;w) + f_{\mathcal{Q}}^k(\tilde{\mathbf{x}}_i^p;w)\nabla_w f_{\mathcal{Q}}^l(\tilde{\mathbf{x}}_i^p;w)\right) \\ & + \frac{1}{\sum_{s\neq j} e^{\tilde{\mathcal{Q}}_j^{q\top}\tilde{\mathcal{Q}}_s^q}}\sum_{k\neq l} e^{\tilde{\mathcal{Q}}_l^{p\top}\tilde{\mathcal{Q}}_k^p}\left(f_{\mathcal{Q}}^l(\tilde{\mathbf{x}}_i^q;w)\nabla_w f_{\mathcal{Q}}^k(\tilde{\mathbf{x}}_i^q;w) + f_{\mathcal{Q}}^k(\tilde{\mathbf{x}}_i^q;w)\nabla_w f_{\mathcal{Q}}^l(\tilde{\mathbf{x}}_i^q;w)\right)\Bigg] \\ & + \sum_{p=1}^{m}\sum_{l=1}^{K}\left[\log\bar{\tilde{\mathcal{Q}}}_l^p + 1\right]\left(\nabla_w f_{\mathcal{Q}}^l(\tilde{\mathbf{x}}_i^p)\right). \end{aligned}$$

Since $\mathcal{L}(z,z';\phi,w)$ is second order differentiable, the gradient of $\mathcal{L}(z,z';\phi,w)$ w.r.t. ϕ exists, which is formulated as

$$\nabla_\phi \mathcal{L}(z_i;\hat{w}(\phi))\Big|_{\phi^{(t)}} = -\frac{\eta_w}{n_e}\sum_{s=1}^{n_e} G_{is}\nabla_\phi g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}} - \frac{\eta_w}{n_e(n_e-1)}\sum_{s=1}^{n_e}\sum_{k\neq s} G_{isk}\nabla_\phi g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}},$$

with

$$\begin{aligned}G_{is} :=& \left(\nabla_{\hat{w}}\ell^{(1)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}^{\top} + \nabla_{\hat{w}}\ell^{(2)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}^{\top}\right)\nabla_w\ell^{(1)}(\tilde{z}_s;w)\Big|_{\hat{w}^{(t)}}\\ &+\left(\nabla_{\hat{w}}\ell^{(1)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}^{\top} + \nabla_{\hat{w}}\ell^{(2)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}^{\top}\right)\nabla_w\ell^{(2)}(\tilde{z}_s;w)\Big|_{\hat{w}^{(t)}}\\ G_{isk} :=& \left(\nabla_{\hat{w}}\ell^{(1)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}^{\top} + \nabla_{\hat{w}}\ell^{(2)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}^{\top}\right)\nabla_w\ell(\tilde{z}_s,\tilde{z}_k;w)\Big|_{\hat{w}^{(t)}},\end{aligned}$$

and

$$\nabla_\phi\mathcal{L}(z_i,z_j;\hat{w}(\phi))\Big|_{\phi^{(t)}} = -\frac{\eta_w}{n_e}\sum_{s=1}^{n_e}G_{ijs}\nabla_\phi g(\tilde{z}_s;\phi) - \frac{\eta_w}{n_e(n_e-1)}\sum_{s=1}^{n_e}\sum_{k\neq s}G_{ijsk}\nabla_\phi g(\tilde{z}_s;\phi),$$

with

$$\begin{aligned}G_{ijs} :=& \nabla_{\hat{w}}\ell(z_i,z_j;\hat{w})\Big|_{\hat{w}^{(t)}}^{\top}\left(\nabla_w\ell^{(1)}(\tilde{z}_s;w)\Big|_{\hat{w}^{(t)}} + \nabla_w\ell^{(2)}(\tilde{z}_s;w)\Big|_{\hat{w}^{(t)}}\right)\\ G_{ijsk} :=& \nabla_{\hat{w}}\ell(z_i,z_j;\hat{w})\Big|_{\hat{w}^{(t)}}^{\top}\nabla_w\ell(\tilde{z}_s,\tilde{z}_k;w)\Big|_{\hat{w}^{(t)}}.\end{aligned}$$

Since $\mathcal{L}(z,z';\phi,w)$ is second order differentiable with bounded gradient, there exists a constant $\rho_0 < \infty$ such that $\|\nabla_w\ell^{(1)}(z;w)\| \leq \rho_0$, $\|\nabla_w\ell^{(2)}(z;w)\| \leq \rho_0$ and $\|\nabla_w\ell(z,z';w)\| \leq \rho_0$ hold. Since $g(\tilde{z};\phi)$ is second order differentiable with bounded gradient, there exists constants $\rho_2 < \infty$ such that $\|\nabla_\phi g(\tilde{z};\phi)\| \leq \rho_1$ hold. This implies that $|G_{is}| \leq 4\rho_0^2$, $|G_{ijs}|, |G_{isk}| \leq 2\rho_0^2$ and $|G_{ijsk}| \leq \rho_0^2$ hold. Thus, we have $\|\nabla_\phi\mathcal{L}_i(\hat{w}(\phi))\big|_{\phi^{(t)}}\| \leq 6\eta_w\rho_1\rho_0^2$ and $\|\nabla_\phi\mathcal{L}_{ij}(\hat{w}(\phi))\big|_{\phi^{(t)}}\| \leq 3\eta_w\rho_1\rho_0^1$. Thus, the upper-level objective $\mathcal{L}(z,z';\phi,w)$ as a function of ϕ is $L = 6\rho_1\rho_0^2$-Lipschitz continuous. Further, the gradient of $\nabla_\phi\mathcal{L}(z_i;\hat{w}(\phi))\big|_{\phi^{(t)}}$ and $\nabla_\phi\mathcal{L}(z_i,z_j;\hat{w}(\phi))\big|_{\phi^{(t)}}$ w.r.t. ϕ are formulated as

$$\begin{aligned}&\nabla^2_{\phi^2}\mathcal{L}(z_i;\hat{w}(\phi))\Big|_{\phi^{(t)}}\\ =&-\frac{\eta_w}{n_e}\sum_{s=1}^{n_e}\left(\nabla_\phi g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}}\nabla_\phi G_{is}\Big|_{\phi^{(t)}}^{\top} + G_{is}\nabla^2_{\phi^2}g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}}^{\top}\right)\\ &-\frac{\eta_w}{n_e(n_e-1)}\sum_{s=1}^{n_e}\sum_{k\neq s}\left(\nabla_\phi g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}}\nabla_\phi G_{isk}\Big|_{\phi^{(t)}}^{\top} + G_{isk}\nabla^2_{\phi^2}g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}}^{\top}\right),\end{aligned}$$

and

$$\begin{aligned}&\nabla^2_{\phi^2}\mathcal{L}(z_i,z_j;\hat{w}(\phi))\Big|_{\phi^{(t)}}\\ =&-\frac{\eta_w}{n_e}\sum_{s=1}^{n_e}\left(\nabla_\phi g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}}\nabla_\phi G_{ijs}\Big|_{\phi^{(t)}}^{\top} + G_{ijs}\nabla^2_{\phi^2}g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}}^{\top}\right)\\ &-\frac{\eta_w}{n_e(n_e-1)}\sum_{s=1}^{n_e}\sum_{k\neq s}\left(\nabla_\phi g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}}\nabla_\phi G_{ijsk}\Big|_{\phi^{(t)}}^{\top} + G_{ijsk}\nabla^2_{\phi^2}g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}}^{\top}\right),\end{aligned}$$

where

$$\begin{aligned}\nabla_\phi G_{is} =& \left(\nabla^2_{\hat{w}^2}\ell^{(1)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}F + \nabla^2_{\hat{w}^2}\ell^{(2)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}F\right)^{\top}\nabla_w\ell^{(1)}(\tilde{z}_s;w)\Big|_{\hat{w}^{(t)}}\\ &+\left(\nabla^2_{\hat{w}^2}\ell^{(1)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}F + \nabla^2_{\hat{w}^2}\ell^{(2)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}F\right)^{\top}\nabla_w\ell^{(2)}(\tilde{z}_s;w)\Big|_{\hat{w}^{(t)}}\\ \nabla_\phi G_{isk} =& \left(\nabla^2_{\hat{w}^2}\ell^{(1)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}F + \nabla^2_{\hat{w}^2}\ell^{(2)}(z_i;\hat{w})\Big|_{\hat{w}^{(t)}}F\right)^{\top}\nabla_w\ell(\tilde{z}_s,\tilde{z}_k;w)\Big|_{\hat{w}^{(t)}}\\ \nabla_\phi G_{ijs} =& \left(\nabla^2_{\hat{w}^2}\ell(z_i,z_j;\hat{w})\Big|_{\hat{w}^{(t)}}F\right)^{\top}\left(\nabla_w\ell^{(1)}(\tilde{z}_s;w)\Big|_{\hat{w}^{(t)}} + \nabla_w\ell^{(2)}(\tilde{z}_s;w)\Big|_{\hat{w}^{(t)}}\right)\\ \nabla_\phi G_{ijsk} =& \left(\nabla^2_{\hat{w}^2}\ell(z_i,z_j;\hat{w})\Big|_{\hat{w}^{(t)}}F\right)^{\top}\nabla_w\ell(\tilde{z}_s,\tilde{z}_k;w)\Big|_{\hat{w}^{(t)}},\end{aligned}$$

with

$$\begin{aligned}F :=& -\frac{\eta_w}{n_e}\sum_{s=1}^{n_e}\left(\nabla_w\ell^{(1)}(\tilde{z}_s;w)\Big|_{\hat{w}^{(t)}}\nabla_\phi g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}}^{\top} + \nabla_w\ell^{(2)}(\tilde{z}_s;w)\Big|_{\hat{w}^{(t)}}\nabla_\phi g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}}^{\top}\right)\\ &-\frac{\eta_w}{n_e(n_e-1)}\sum_{s=1}^{n_e}\sum_{k\neq s}\nabla_w\ell(\tilde{z}_s,\tilde{z}_k;w)\Big|_{\hat{w}^{(t)}}\nabla_\phi g(\tilde{z}_s;\phi)\Big|_{\phi^{(t)}}^{\top}.\end{aligned}$$

Since $\mathcal{L}(z, z'; \phi, w)$ is second order differentiable with bounded Hessian, there exists a constant $\rho_2 < \infty$ such that $\|\nabla^2_{\hat{w}^2}\ell^{(1)}(z_i; \hat{w})\| \leq \rho_2$, $\|\nabla^2_{\hat{w}^2}\ell^{(2)}(z_i; \hat{w})\| \leq \rho_2$ and $\|\nabla^2_{\hat{w}^2}\ell(z_i, z_j; \hat{w})\| \leq \rho_2$ hold. Since $g(\tilde{z}; \phi)$ is second order differentiable with bounded Hessian, there exists constants $\rho_3 < \infty$ such that $\|\nabla^2_{\phi^2} g(\tilde{z}; \phi)\| \leq \rho_3$ hold. Thus, we have $\|F\| \leq 3\eta_w\rho_0\rho_1$. This implies that $\|\nabla_\phi G_{is}\| \leq 12\eta_w\rho_0^2\rho_1\rho_2$, $\|\nabla_\phi G_{isk}\|, \|\nabla_\phi G_{ijs}\| \leq 6\eta_w\rho_0^2\rho_1\rho_2$ and $\|\nabla_\phi G_{ijsk}\| \leq 3\eta_w\rho_0^2\rho_1\rho_2$ holds. Then we have

$$\begin{aligned}\|\nabla^2_{\phi^2}\mathcal{L}_i(\hat{w}(\phi))\big|_{\phi^{(t)}}\| \leq& 6\eta_w^2\rho_0^2(3\rho_1^2\rho_2 + \rho_3),\\ \|\nabla^2_{\phi^2}\mathcal{L}_{ij}(\hat{w}(\phi))\big|_{\phi^{(t)}}\| \leq& 3\eta_w^2\rho_0^2(3\rho_1^2\rho_2 + \rho_3).\end{aligned}$$

Thus, the upper-level objective $\mathcal{L}(z, z'; \phi, w)$ as a function of ϕ is $G = 6\eta_w^2\rho_0^2(3\rho_1^2\rho_2 + \rho_3)$-Lipschitz smooth. Next, the update rule of ϕ is formulated as

$$\phi^{(t+1)} = \phi^{(t)} - \eta_\phi(\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)})) + \zeta),$$

where $\zeta = \nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)}))|_{\mathcal{B}} - \nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)}))$ and $\mathcal{B}$ is a mini-batch data sampled i.i.d. from $\mathcal{D}^c$. This indicates that $\mathbb{E}[\zeta] = 0$ holds. First, we have

$$\begin{aligned}&\mathcal{L}(\hat{w}^{(t+1)}(\phi^{(t+1)})) - \mathcal{L}(\hat{w}^{(t)}(\phi^{(t)}))\\ =&\Big[\mathcal{L}(\hat{w}^{(t+1))}(\phi^{(t+1)})) - \mathcal{L}(\hat{w}^{(t)}(\phi^{(t+1)}))\Big] + \Big[\mathcal{L}(\hat{w}^{(t)}(\phi^{(t+1)})) - \mathcal{L}(\hat{w}^{(t)}(\phi^{(t)}))\Big]\\ \leq&\langle\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t+1)})), \hat{w}^{(t+1)}(\phi^{(t+1)}) - \hat{w}^{(t)}(\phi^{(t+1)})\rangle + \frac{\rho_2}{2}\|\hat{w}^{(t+1)}(\phi^{(t+1)}) - \hat{w}^{(t)}(\phi^{(t+1)})\|_2^2\\ &+ \langle\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)})), \phi^{(t+1)} - \phi^{(t)}\rangle + \frac{G}{2}\|\phi^{(t+1)} - \phi^{(t)}\|_2^2.\end{aligned}$$

According to the definition of $\hat{w}$, one can verify that

$$\langle\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t+1)})), \hat{w}^{(t+1)}(\phi^{(t+1)}) - \hat{w}^{(t)}(\phi^{(t+1)})\rangle + \frac{\rho_2}{2}\|\hat{w}^{(t+1)}(\phi^{(t+1)}) - \hat{w}^{(t)}(\phi^{(t+1)})\|_2^2 \leq 18\eta_w\rho_0^2 + 18\rho_2\eta_w^2\rho_0^2.$$

Besides, we have

$$\begin{aligned}&\langle\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)})), \phi^{(t+1)} - \phi^{(t)}\rangle + \frac{G}{2}\|\phi^{(t+1)} - \phi^{(t)}\|_2^2\\ =&\langle\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)})), -\eta_\phi[\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)})) + \zeta]\rangle + \frac{G\eta_\phi^2}{2}\|\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)})) + \zeta\|_2^2\\ =&-(\eta_\phi - \frac{G\eta_\phi^2}{2})\|\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)}))\|_2^2 + \frac{G\eta_\phi^2}{2}\|\zeta\|_2^2 - (\eta_\phi - G\eta_\phi^2)\langle\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)})), \zeta\rangle.\end{aligned}$$

Therefore,

$$\begin{aligned}&(\eta_\phi - \frac{G\eta_\phi^2}{2})\|\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)}))\|_2^2\\ \leq&\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)})) - \mathcal{L}(\hat{w}^{(t+1)}(\phi^{(t+1)})) + 18\eta_w\rho_0^2 + 18\rho_2\eta_w^2\rho_0^2 + \frac{G\eta_\phi^2}{2}\|\zeta\|_2^2 - (\eta_\phi - G\eta_\phi^2)\langle\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)})), \zeta\rangle.\end{aligned}$$

Taking summation on both sides, we have

$$\begin{aligned}&\sum_{t=1}^{T}(\eta_t - \frac{G\eta_t^2}{2})\|\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)}))\|_2^2\\ \leq&\mathcal{L}(\hat{w}^{(1)}(\phi^{(1)})) - \mathcal{L}(\hat{w}^{(T)}(\phi^{(T)})) + 18\eta_w\rho_0^2T(1 + \eta_w\rho_2) + \frac{G\eta_\phi^2}{2}\sum_{t=1}^{T}\|\zeta\|_2^2 - \sum_{t=1}^{T}(\eta_t - G\eta_t^2)\langle\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)})), \zeta\rangle.\end{aligned}$$

Further, taking the expectation with respect to ζ on both sides, one can find that

$$\begin{aligned}&\sum_{t=1}^{T}(\eta_\phi - \frac{G\eta_\phi^2}{2})\mathbb{E}_\zeta\|\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)}))\|^2\\ \leq&\mathcal{L}(\hat{w}^{(1)}(\phi^{(1)})) - \mathcal{L}(\hat{w}^{(T)}(\phi^{(T)})) + 18\eta_w\rho_0^2T(1 + \eta_w\rho_2) + \frac{G\eta_\phi^2}{2}\sum_{t=1}^{T}\|\zeta\|^2\\ \leq&\mathcal{L}(\hat{w}^{(1)}(\phi^{(1)})) - \mathcal{L}(\hat{w}^{(T)}(\phi^{(T)})) + 18\eta_w\rho_0^2T(1 + \eta_w\rho_2) + \frac{G\eta_\phi^2T\sigma^2}{2}.\end{aligned}$$

Thus we have

$$
\begin{aligned}
&\min_t \mathbb{E}\|\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)}))\|^2 \leq \frac{\sum_{t=1}^{T}(\eta_\phi - \frac{L\eta_\phi^2}{2})\mathbb{E}_{\zeta^{(B)}}\|\nabla\mathcal{L}(\hat{w}^{(t)}(\phi^{(t)}))\|_2^2}{\sum_{t=1}^{T}(\eta_\phi - \frac{L\eta_\phi^2}{2})} \\
\leq &\frac{2\mathcal{L}(\hat{w}^{(1)}(\phi^{(1)})) - 2\mathcal{L}(\hat{w}^{(T)}(\phi^{(T)})) + 36\eta_w\rho_0^2 T\,(1+\eta_w\rho_2) + G\eta_\phi^2 T\sigma^2}{\sum_{t=1}^{T}(2\eta_\phi - L\eta_\phi^2)} \\
\leq &\frac{2\mathcal{L}(\hat{w}^{(1)}(\phi^{(1)})) - 2\mathcal{L}(\hat{w}^{(T)}(\phi^{(T)})) + 36\eta_w\rho_0^2 T\,(1+\eta_w\rho_2) + G\eta_\phi^2 T\sigma^2}{T\eta_\phi} \\
= &\frac{2\mathcal{L}(\hat{w}^{(1)}(\phi^{(1)}))}{T\eta_\phi} - \frac{2\mathcal{L}(\hat{w}^{(T)}(\phi^{(T)}))}{T\eta_\phi} + \frac{36\eta_w\rho_0^2(1+\eta_w\rho_2)}{\eta_\phi} + G\sigma^2\eta_\phi \\
\leq &\frac{2\mathcal{L}(\hat{w}^{(1)}(\phi^{(1)}))}{T}\max\{L,\frac{\sqrt{T}}{C}\} + \frac{2|\mathcal{L}(\hat{w}^{(T)}(\phi^{(T)}))|}{T}\max\{L,\frac{\sqrt{T}}{C}\} + 36\min\{1,\frac{k}{T}\}\max\{L,\frac{\sqrt{T}}{C}\}\rho_0^2(\eta_w\rho_2+1) \\
&+ G\sigma^2\min\{\frac{1}{L},\frac{C}{\sqrt{T}}\} \\
= &\mathcal{O}(\frac{1}{\sqrt{T}}).
\end{aligned}
$$

A.2. Proof of Theorem 3.3

Proof. According to the definition in the main paper, the objective of bi-level optimization is formulated as

$$
\begin{aligned}
&\min_{\phi,w}\mathcal{L}(f(\mathcal{D}^c;w^*(\phi))) \\
&w^*(\phi) = \operatorname*{argmin}_w \mathcal{L}(f(\mathcal{D}^c;w)) + \mathcal{L}(f(\mathcal{D}^e;w), g(\mathcal{D}^e;\phi)).
\end{aligned}
$$

According to the definition of g, there exists ϕ' such that $g(\mathcal{D}^e;\phi') = 0$ holds. Thus we have $\hat{\mathcal{L}}(w^*) = \hat{\mathcal{L}}(w^*(\phi'))$. Since $\hat{\phi}$ is the optimal solution, for any $\phi \in \Phi$, the following equality holds

$$
\hat{\mathcal{L}}(w^*(\hat{\phi})) \leq \hat{\mathcal{L}}(w^*(\phi)).
$$

Replacing ϕ with ϕ', we have $\hat{\mathcal{L}}(w^*(\hat{\phi})) \leq \hat{\mathcal{L}}(w^*)$. This finishes the proof.

A.3. Proof of Theorem Theorem 3.4

To prove Theorem 3.4, we first introduce the following three lemmas.

Lemma A.1. *Assume that $\|f_{\mathcal{Z}}(\mathbf{x}^p)\|_\infty \leq E$ holds for all $\{\mathbf{x}^p\}_{p=1}^m \in \mathcal{X}$. We define the empirical risk and its expectation as*

$$
\widehat{\mathcal{L}}(f) = \sum_{p=1}^{m}\sum_{q=p+1}^{m}\left[-\frac{2}{n}\sum_{i=1}^{n} f_{\mathcal{Z}}(\mathbf{x}_i^p;w)^\top f_{\mathcal{Z}}(\mathbf{x}_i^q;w) + \frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n}\left(f_{\mathcal{Z}}(\mathbf{x}_i^p;w)^\top f_{\mathcal{Z}}(\mathbf{x}_j^q;w)\right)^2\right],
$$

and

$$
\mathcal{L}(f) = \sum_{p=1}^{m}\sum_{q=p+1}^{m}\mathbb{E}_{\mathbf{x}^p,\mathbf{x}^q}\left[-2f_{\mathcal{Z}}(\mathbf{x}^p;w)^\top f_{\mathcal{Z}}(\mathbf{x}^q;w) + \left(f_{\mathcal{Z}}(\mathbf{x}^p;w)^\top f_{\mathcal{Z}}(\mathbf{x}^q;w)\right)^2\right].
$$

With probability at least $1-\delta$, for any $f \in \mathcal{F}$, we have

$$
\mathcal{L}(f) \leq \widehat{\mathcal{L}}(f) + \frac{4D^2E^4m(m-1)}{\sqrt{n}} + 4D^2E^4m(m-1)\sqrt{\frac{1}{2n}\log\frac{1}{\delta}}.
$$

Proof. This proof is inspired by (Li & Liu, 2021). We first discuss the case where $m=2$. Let $z_i := (\mathbf{x}_i^p, \mathbf{x}_i^q)$ be the multi-view tuple, then the sample set can be denoted as $S := \{z_i\}_{i=1}^n$. For simplicity, The first and the second part of the loss are denoted as $f_{\mathcal{Z}}(z_i) := f_{\mathcal{Z}}(\mathbf{x}_i^p;w)^\top f_{\mathcal{Z}}(\mathbf{x}_i^q;w)$ and $f_{\mathcal{Z}}(z_i,z_j) := [(f(\mathbf{x}_i^p;w)^\top f(\mathbf{x}_j^q;w))^2 + (f(\mathbf{x}_j^p;w)^\top f(\mathbf{x}_i^q;w))^2]/2$,

respectively. One can verify that $f_{\mathcal{Z}}(z, z')$ is symmetric function, *i.e.*, $f_{\mathcal{Z}}(z, z') = f_{\mathcal{Z}}(z', z)$. Then the empirical risk and its expectation can be formulated as

$$\widehat{\mathcal{L}}(f) = -\frac{2}{n}\sum_{i=1}^{n} f_{\mathcal{Z}}(z_i) + \frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n} f_{\mathcal{Z}}(z_i, z_j),$$

and

$$\mathcal{L}(f) = -2\mathbb{E}_z\left[f_{\mathcal{Z}}(z)\right] + \mathbb{E}_{z,z'}\left[f_{\mathcal{Z}}(z, z')\right].$$

Let $\bar{S}$ be the sample set that different from S by only one tuple $z_r := (\bar{\mathbf{x}}_r^p, \bar{\mathbf{x}}_r^q)$. The empirical risk on $\bar{S}$ is denoted as $\widehat{\mathcal{L}}'_n$. We have

$$\begin{aligned}
&\left|\sup_{f\in\mathcal{F}}|\mathcal{L} - \widehat{\mathcal{L}}(f)| - \sup_{f\in\mathcal{F}}|\mathcal{L}(f) - \widehat{\mathcal{L}}'(f)|\right| \\
\leq &\sup_{f\in\mathcal{F}}|\widehat{\mathcal{L}}(f) - \widehat{\mathcal{L}}'(f)| \\
\leq &\sup_{f\in\mathcal{F}}\left|-\frac{2}{n}\Big(f_{\mathcal{Z}}(z_r) - f_{\mathcal{Z}}(\bar{z}_r)\Big)\right| + \sup_{f\in\mathcal{F}}\left|\frac{1}{n(n-1)}\sum_{i\neq r}\Big(f_{\mathcal{Z}}(z_r, z_i) - f_{\mathcal{Z}}(\bar{z}_r, z_i)\Big) + \Big(f_{\mathcal{Z}}(z_i, z_r) - f_{\mathcal{Z}}(z_i, \bar{z}_r)\Big)\right| \\
\leq &\frac{8D^2E^4}{n}.
\end{aligned}$$

Then we analyze the upper bound of the expectation term, *i.e.*, $\mathbb{E}\sup_{f\in\mathcal{F}}|\widehat{\mathcal{L}}(f) - \mathcal{L}(f)|$. First we have

$$\begin{aligned}
&\mathbb{E}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}|\widehat{\mathcal{L}}(f) - \mathcal{L}_S(f)| \\
=&\mathbb{E}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n} f_{\mathcal{Z}}(z_i) - \mathbb{E}_z\left[f_{\mathcal{Z}}(z)\right]\right| + \mathbb{E}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left|\frac{1}{n(n-1)}\sum_{i\neq j} f_{\mathcal{Z}}(z_i, z_j) - \mathbb{E}_{z,z'}\left[f_{\mathcal{Z}}(z, z')\right]\right| \\
\leq&\mathbb{E}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n} f_{\mathcal{Z}}(z_i) - \mathbb{E}_z\left[f_{\mathcal{Z}}(z)\right]\right| + \mathbb{E}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} f_{\mathcal{Z}}(z_i, z_{i+\lfloor n/2\rfloor}) - \mathbb{E}_{z,z'}\left[f_{\mathcal{Z}}(z, z')\right]\right|,
\end{aligned}$$

where the last inequality is obtained by the Lemma A.1 in (Clémençon et al., 2008). Let $\sigma_1, \ldots, \sigma_n$ be i.i.d. independent random variables taking values in $\{-1, 1\}$ and $\bar{S} := \{\bar{z}, \ldots, \bar{z}_n\}$ be the independent copy of $S = \{z_1, \ldots, z_n\}$, the last term can be bound by

$$\begin{aligned}
&\mathbb{E}_{S,\bar{S}}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}[f_{\mathcal{Z}}(z_i) - f_{\mathcal{Z}}(\bar{z}_i)]\right| + \mathbb{E}_{S,\bar{S},\boldsymbol{\sigma}}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} f_{\mathcal{Z}}(z_i, z_{i+\lfloor n/2\rfloor}) - \frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} f_{\mathcal{Z}}(\bar{z}_i, \bar{z}_{i+\lfloor n/2\rfloor})\right| \\
=&\mathbb{E}_{S,\bar{S},\boldsymbol{\sigma}}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i[f_{\mathcal{Z}}(z_i) - f_{\mathcal{Z}}(\bar{z}_i)]\right| + \mathbb{E}_{S,\bar{S},\boldsymbol{\sigma}}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\sigma_i[f_{\mathcal{Z}}(z_i, z_{i+\lfloor n/2\rfloor}) - f_{\mathcal{Z}}(\bar{z}_i, \bar{z}_{i+\lfloor n/2\rfloor})]\right| \\
=&4\mathbb{E}_{S,\boldsymbol{\sigma}}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{Z}}(z_i)\right| + 2\mathbb{E}_{S,\boldsymbol{\sigma}}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\sigma_i f_{\mathcal{Z}}(z_i, z_{i+\lfloor n/2\rfloor})\right| \\
\leq&4\mathbb{E}_{S,\boldsymbol{\sigma}}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}[f_{\mathcal{Z}}(z_i)]^2\right)^{\frac{1}{2}} + 2\mathbb{E}_{S,\boldsymbol{\sigma}}\sup_{f_{\mathcal{Z}}\in\mathcal{F}}\left(\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}[f_{\mathcal{Z}}(z_i, z_{i+\lfloor n/2\rfloor})]^2\right)^{\frac{1}{2}} \\
\leq&\frac{8D^2E^4}{\sqrt{n}},
\end{aligned}$$

where the last inequality is obtain by the Khintchine-Kahane inequality (Latała & Oleszkiewicz, 1994). Thus, according to the McDiarmid inequality (Mohri et al., 2018), with probability at least $1-\delta$ for any $f \in \mathcal{F}$, we have

$$\mathcal{L}(f) \leq \widehat{\mathcal{L}}(f) + \frac{8D^2E^4}{\sqrt{n}} + 8D^2E^4\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

The case where $m > 2$, *i.e.*, That is, with probability at least $1-\delta$ for any $f \in \mathcal{F}$, we have

$$\mathcal{L}(f) \leq \widehat{\mathcal{L}}(f) + \frac{4D^2E^4m(m-1)}{\sqrt{n}} + 4D^2E^4m(m-1)\sqrt{\frac{1}{2n}\log\frac{1}{\delta}}.$$

Lemma A.2. *We define the empirical risk and its expectation as*

$$\begin{aligned}\widehat{\mathcal{L}}(f) =& -\frac{1}{nK}\sum_{p=1}^{m}\sum_{q=p+1}^{m}\sum_{l=1}^{K}\log\frac{e^{\mathcal{Q}_l^{p\top}\mathcal{Q}_l^q}}{\sum_{s\neq l}e^{\mathcal{Q}_l^{p\top}\mathcal{Q}_s^p}}\\ =&\frac{1}{nK}\sum_{p=1}^{m}\sum_{q=p+1}^{m}\sum_{l=1}^{K}\left[\log\left(\sum_{s\neq l}\exp\left\{\sum_{i=1}^{n}f_{\mathcal{Q}}^l(\mathbf{x}_i^p))f_{\mathcal{Q}}^s(\mathbf{x}_i^p)\right\}\right)-\sum_{i=1}^{n}f_{\mathcal{Q}}^l(\mathbf{x}_i^p)f_{\mathcal{Q}}^l(\mathbf{x}_i^q)\right],\end{aligned}$$

and

$$\mathcal{L}(f)=\frac{1}{K}\sum_{p=1}^{m}\sum_{q=p+1}^{m}\sum_{l=1}^{K}\left[\log\left(\sum_{s\neq l}\exp\left\{\mathbb{E}_{\mathbf{x}^p,\mathbf{x}^q}\left[f_{\mathcal{Q}}^l(\mathbf{x}^p)f_{\mathcal{Q}}^s(\mathbf{x}^p)\right]\right\}\right)-\mathbb{E}_{\mathbf{x}^p,\mathbf{x}^q}\left[f_{\mathcal{Q}}^l(\mathbf{x}^p)f_{\mathcal{Q}}^l(\mathbf{x}^q)\right]\right].$$

With probability at least $1-\delta$, the following inequality holds

$$\mathcal{L}(f)\leq\widehat{\mathcal{L}}(f)+\frac{m(m-1)(K+2)}{2\sqrt{n}}+\frac{m(m-1)(K+1)}{2}\sqrt{\frac{1}{2n}\log\frac{1}{\delta}}.$$

Proof. We first discuss the case when $m=2$. Let $z_i:=(\mathbf{x}_i^p,\mathbf{x}_i^q)$ be the multi-view tuple, then the sample set can be denoted as $S:=\{z_i\}_{i=1}^n$. For simplicity, the first and the second part of the loss are denoted as $f_{\mathcal{Q}}^{l,s}(z_i):=f_{\mathcal{Q}}^l(\mathbf{x}_i^p;w)^\top f_{\mathcal{Q}}^s(\mathbf{x}_i^p;w)$ and $f_{\mathcal{Q}}^l(z_i):=f_{\mathcal{Q}}^l(\mathbf{x}_i^p)f_{\mathcal{Q}}^l(\mathbf{x}_i^q)$, respectively. Let $\mathcal{F}_{\mathcal{Q}}^{l,s}$ and $\mathcal{F}_{\mathcal{Q}}^l$ be the family of $f_{\mathcal{Q}}^{l,s}(z_i)$ and $f_{\mathcal{Q}}^l(z_i)$, respectively. Then the empirical risk and its expectation can be formulated as

$$\widehat{\mathcal{L}}(f)=\frac{1}{nK}\sum_{l=1}^{K}\log\left(\sum_{s\neq l}\exp\left\{\sum_{i=1}^{n}f_{\mathcal{Q}}^{l,s}(z_i)\right\}\right)-\frac{1}{nK}\sum_{l=1}^{K}\sum_{i=1}^{n}f_{\mathcal{Q}}^l(z_i):=\widehat{\mathcal{L}}^{(1)}(f)+\widehat{\mathcal{L}}^{(2)}(f),$$

and

$$\mathcal{L}(f)=\frac{1}{K}\sum_{l=1}^{K}\log\left(\sum_{s\neq l}\exp\left\{\mathbb{E}_z\left[f_{\mathcal{Q}}^{l,s}(z)\right]\right\}\right)-\frac{1}{K}\sum_{l=1}^{K}\mathbb{E}_z\left[f_{\mathcal{Q}}^l(z)\right]:=\mathcal{L}^{(1)}(f)+\mathcal{L}^{(1)}(f).$$

We introduce the following empirical risk $\hat{L}(f)$ and its expectation $L(f)$ to change the log-sum-exp term into a more concise form,

$$\begin{aligned}\hat{L}(f)&=\frac{1}{nK(K-1)}\sum_{l=1}^{K}\sum_{s\neq l}\sum_{i=1}^{n}f_{\mathcal{Q}}^{l,s}(z_i),\\ L(f)&=\frac{1}{K(K-1)}\sum_{l=1}^{K}\sum_{s\neq l}^{K}\mathbb{E}_z\left[f_{\mathcal{Q}}^{l,s}(z)\right].\end{aligned}$$

One can verify that

$$\begin{aligned}\hat{L}(f)\leq\widehat{\mathcal{L}}^{(1)}(f)-\frac{\log(K-1)}{n}\leq(K-1)\hat{L}(f),\\ L(f)\leq\mathcal{L}^{(1)}(f)-\frac{\log(K-1)}{n}\leq(K-1)L(f),\end{aligned}$$

holds. Then we have

$$\begin{aligned}|\mathcal{L}^{(1)}(f)-\widehat{\mathcal{L}}^{(1)}|=&\left|\left(\mathcal{L}^{(1)}(f)-\frac{\log(K-1)}{n}\right)-\left(\widehat{\mathcal{L}}^{(1)}-\frac{\log(K-1)}{n})\right)\right|\\ \leq&\max\{|(K-1)L(f)-\hat{L}(f)|,|L(f)-(K-1)\hat{L}(f)|\}.\end{aligned}$$

Without loss of generality, we assume that $\max\{|(K-1)L(f)-\hat{L}(f)|,|L(f)-(K-1)\hat{L}(f)|\}=|(K-1)L(f)-\hat{L}(f)|$. Let $\bar{S}$ be the sample set that different from S by only one tuple $z_r:=(\bar{\mathbf{x}}_r^p,\bar{\mathbf{x}}_r^q)$. The empirical risk on $\bar{S}$ is denoted as $\widehat{\mathcal{L}}'$. We have

$$\begin{aligned}&\left|\sup_{f\in\mathcal{F}}|(K-1)L(f)-\hat{L}(f)|-\sup_{f\in\mathcal{F}}|(K-1)L(f)-\hat{L}'(f)|\right|\\ \leq&\sup_{f\in\mathcal{F}}|\hat{L}(f)-\hat{L}'(f)|\\ \leq&\sup_{f\in\mathcal{F}}\left|\frac{1}{nK(K-1)}\sum_{l=1}^{K}\sum_{s\neq l}\left(f_{\mathcal{Q}}^{l,s}(z_r)-f_{\mathcal{Q}}^{l,s}(\bar{z}_r)\right)\right|\\ \leq&\frac{1}{n}.\end{aligned}$$

Then we analyze the upper bound of the expectation term, *i.e.*, $\mathbb{E}\sup_{f\in\mathcal{F}}|\widehat{\mathcal{L}}(f)-\mathcal{L}(f)|$. Let $\sigma_1,\ldots,\sigma_n$ be i.i.d. independent random variables taking values in $\{-1,1\}$ and $\bar{S}:=\{\bar{z}_1,\ldots,\bar{z}_n\}$ be the independent copy of $S=\{z_1,\ldots,z_n\}$, we have

$$
\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}}|(K-1)L(f)-\hat{L}(f)|\\
\leq&\mathbb{E}_{S,\bar{S}}\sup_{f\in\mathcal{F}}\left|\frac{1}{nK(K-1)}\sum_{l=1}^{K}\sum_{s\neq l}\sum_{i=1}^{n}\left(f_{\mathcal{Q}}^{l,s}(z_i)-f_{\mathcal{Q}}^{l,s}(\bar{z}_i)\right)\right|+(K-2)\mathbb{E}_{\bar{S}}\sup_{f\in\mathcal{F}}\left|\frac{1}{nK(K-1)}\sum_{l=1}^{K}\sum_{s\neq l}\sum_{i=1}^{n}f_{\mathcal{Q}}^{l,s}(\bar{z}_i)\right|\\
\leq&\mathbb{E}_{S,\bar{S},\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\left|\frac{1}{nK(K-1)}\sum_{l=1}^{K}\sum_{s\neq l}\sum_{i=1}^{n}\sigma_i\left(f_{\mathcal{Q}}^{l,s}(z_i)-f_{\mathcal{Q}}^{l,s}(\bar{z}_i)\right)\right|+(K-2)\mathbb{E}_{\bar{S},\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\left|\frac{1}{nK(K-1)}\sum_{l=1}^{K}\sum_{s\neq l}\sum_{i=1}^{n}\sigma_i f_{\mathcal{Q}}^{l,s}(\bar{z}_i)\right|\\
=&K\mathbb{E}_{S,\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\left|\frac{1}{nK(K-1)}\sum_{l=1}^{K}\sum_{s\neq l}\sum_{i=1}^{n}\sigma_i f_{\mathcal{Q}}^{l,s}(z_i)\right|\\
\leq&K\max_{l,s}\mathbb{E}_{S,\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{Q}}^{l,s}(z_i)\right|=K\mathcal{R}_n(\mathcal{F}_{\mathcal{Q}}^{l,s}).
\end{aligned}
$$

According to the Khintchine-Kahane inequality (Latała & Oleszkiewicz, 1994), this term is bound by

$$
\mathcal{R}_n(\mathcal{F}_{\mathcal{Q}}^{l,s})=\mathbb{E}_{S,\boldsymbol{\sigma}}\sup_{f_{\mathcal{Q}}^{l,s}\in\mathcal{F}_{\mathcal{Q}}^{l,s}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_{\mathcal{Q}}^{l,s}(z_i)\right|\leq\mathbb{E}_{S,\boldsymbol{\sigma}}\sup_{f_{\mathcal{Q}}^{l,s}\in\mathcal{F}_{\mathcal{Q}}^{l,s}}\frac{1}{n}\left(\sum_{i=1}^{n}[f_{\mathcal{Q}}^{l,s}(z_i)]^2\right)^{\frac{1}{2}}\leq\frac{1}{\sqrt{n}}.
$$

Thus, according to the McDiarmid inequality (Mohri et al., 2018), with probability at least $1-\delta/2$ for any $f\in\mathcal{F}$, we have

$$
\sup_{f\in\mathcal{F}}|(K-1)L(f)-\hat{L}(f))|\leq\frac{K}{\sqrt{n}}+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.
$$

By the same technique, with probability at least $1-\delta/2$ for any $f\in\mathcal{F}$, we have

$$
\sup_{f\in\mathcal{F}}|L(f)-(K-1)\hat{L}(f))|\leq\frac{K}{\sqrt{n}}+K\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.
$$

Note that we have shown $|\mathcal{L}^{(1)}(f)-\widehat{\mathcal{L}}_n^{(1)}|\leq\max\{|(K-1)L(f)-\widehat{L}_n(f)|,|L(f)-(K-1)\widehat{L}_n(f)|$ holds. Thus, with probability at least $1-\delta/2$ for any $f\in\mathcal{F}$, we have

$$
\mathcal{L}^{(1)}(f)\leq\widehat{\mathcal{L}}^{(1)}(f)+\frac{K}{\sqrt{n}}+K\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.
$$

According to the derivation of the generalization bound based on Rademacher complexity (Mohri et al., 2018), with probability $1-\delta/2$ for any $f\in\mathcal{F}$, we have

$$
\mathcal{L}^{(2)}(f)\leq\widehat{\mathcal{L}}^{(2)}(f)+\frac{2}{\sqrt{n}}+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.
$$

Therefore, with probability $1-\delta$ for any $f\in\mathcal{F}$, we have

$$
\mathcal{L}(f)\leq\widehat{\mathcal{L}}(f)+\frac{K+2}{\sqrt{n}}+(K+1)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.
$$

The case where $m>2$ is similar, *i.e.*, with probability at least $1-\delta$ for any $f\in\mathcal{F}$, we have

$$
\mathcal{L}(f)\leq\widehat{\mathcal{L}}(f)+\frac{m(m-1)(K+2)}{2\sqrt{n}}+\frac{m(m-1)(K+1)}{2}\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}.
$$

Lemma A.3. *We define the empirical risk and its expectation as*

$$\widehat{\mathcal{L}}(f)=\sum_{p=1}^{m}\sum_{l=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n}f_{\mathcal{Q}}^{l}(\mathbf{x}_i^p)\right)\log\left(\frac{1}{n}\sum_{j=1}^{n}f_{\mathcal{Q}}^{l}(\mathbf{x}_j^p)\right),$$

and

$$\mathcal{L}(f)=\sum_{p=1}^{m}\sum_{l=1}^{K}\left(\mathbb{E}_{\mathbf{x}}f_{\mathcal{Q}}^{l}(\mathbf{x}^p)\right)\log\left(\mathbb{E}_{\mathbf{x}}f_{\mathcal{Q}}^{l}(\mathbf{x}^p)\right).$$

With probability at least $1-\delta$, the following inequality holds

$$\widehat{\mathcal{L}}(f)\leq\mathcal{L}(f)+\frac{m(m-1)C}{\sqrt{n}}+\frac{m(m-1)C}{2}\sqrt{\frac{\log\frac{2}{\delta}}{2n}},$$

where C is a bounded constant.

Proof. We first discuss the case where $m=1$. Let $f_{\mathcal{Q}}^{j}(\mathbf{x}^p):\mathcal{X}\mapsto\mathbb{R}$ denote the function such that $f_{\mathcal{Q}}^{l}(\mathbf{x}^p)$ is the j-th dimension of $f_{\mathcal{Q}}(\mathbf{x}^p)$, and $\mathcal{F}_{\mathcal{Q}}^{j}$ be the family of $f_{\mathcal{Q}}^{l}$. Define $g(x)=x\log x$, according to the Lagrange Mean Theorem, there exists constant ξ such that $|g(x)-g(y)|\leq|\log\xi+1||x-y|$. According to the deviation of generalization bound based on Rademacher complexity (Mohri et al., 2018), with probability $1-\delta$ for all $f\in\mathcal{F}$, we have

$$\widehat{\mathcal{L}}(f)\leq\mathcal{L}(f)+\frac{2C}{\sqrt{n}}+C\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

where C is a bounded constant. The case where $m\geq2$ is similar. That is, with probability $1-\delta$ for all $f\in\mathcal{F}$, we have

$$\widehat{\mathcal{L}}(f)\leq\mathcal{L}(f)+\frac{m(m-1)C}{\sqrt{n}}+\frac{m(m-1)C}{2}\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Now we give the proof of Theorem 3.4.

Proof. We define the empirical risk and its expectation as

$$\begin{aligned}\widehat{\mathcal{L}}(f(\mathcal{D}^c;w(\phi)))=&\sum_{p=1}^{m}\sum_{q=p+1}^{m}\left[-\frac{2}{n}\sum_{i=1}^{n}f_{\mathcal{Z}}(\mathbf{x}_i^p;w)^{\top}f_{\mathcal{Z}}(\mathbf{x}_i^q;w)+\frac{1}{n(n-1)}\sum_{i,j=1,i\neq j}^{n}\left(f_{\mathcal{Z}}(\mathbf{x}_i^p;w)^{\top}f_{\mathcal{Z}}(\mathbf{x}_j^q;w)\right)^2\right]\\&+\frac{1}{nK}\sum_{p=1}^{m}\sum_{q=p+1}^{m}\sum_{l=1}^{K}\left[\log\left(\sum_{s\neq l}\exp\left\{\sum_{i=1}^{n}f_{\mathcal{Q}}^{l}(\mathbf{x}_i^p))f_{\mathcal{Q}}^{s}(\mathbf{x}_i^p)\right\}\right)-\sum_{i=1}^{n}f_{\mathcal{Q}}^{l}(\mathbf{x}_i^p)f_{\mathcal{Q}}^{l}(\mathbf{x}_i^q)\right]\\&+\sum_{p=1}^{m}\sum_{l=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n}f_{\mathcal{Q}}^{l}(\mathbf{x}_i^p)\right)\log\left(\frac{1}{n}\sum_{j=1}^{n}f_{\mathcal{Q}}^{l}(\mathbf{x}_j^p)\right),\end{aligned}$$

and

$$\begin{aligned}\mathcal{L}(f(w(\phi)))=&\sum_{p=1}^{m}\sum_{q=p+1}^{m}\mathbb{E}_{\mathbf{x}^p,\mathbf{x}^q}\left[-2f_{\mathcal{Z}}(\mathbf{x}^p;w)^{\top}f_{\mathcal{Z}}(\mathbf{x}^q;w)+\left(f_{\mathcal{Z}}(\mathbf{x}^p;w)^{\top}f_{\mathcal{Z}}(\mathbf{x}^q;w)\right)^2\right]\\&+\frac{1}{K}\sum_{p=1}^{m}\sum_{q=p+1}^{m}\sum_{l=1}^{K}\left[\log\left(\sum_{s\neq l}\exp\left\{\mathbb{E}_{\mathbf{x}^p,\mathbf{x}^q}f_{\mathcal{Q}}^{l}(\mathbf{x}^p)f_{\mathcal{Q}}^{s}(\mathbf{x}^p)\right\}\right)-\mathbb{E}_{\mathbf{x}^p,\mathbf{x}^q}f_{\mathcal{Q}}^{l}(\mathbf{x}^p)f_{\mathcal{Q}}^{l}(\mathbf{x}^q)\right]\\&+\sum_{p=1}^{m}\sum_{l=1}^{K}\left(\mathbb{E}_{\mathbf{x}}f_{\mathcal{Q}}^{l}(\mathbf{x}^p)\right)\log\left(\mathbb{E}_{\mathbf{x}}f_{\mathcal{Q}}^{l}(\mathbf{x}^p)\right).\end{aligned}$$

According to Lemma A.1, Lemma A.2 and Lemma A.3, with probability at least $1-\delta$ for any $f\in\mathcal{F}$, we have

$$\mathcal{L}(f(w(\phi)))\leq\widehat{\mathcal{L}}(f(\mathcal{D}^c;w(\phi)))+\frac{\tilde{c}_1}{\sqrt{n}}+\tilde{c}_2\sqrt{\frac{1}{2n}\log\frac{6}{\delta}}.$$

Table 2. Purity comparison (mean±std) of different methods on all benchmark datasets with different missing ratios. Best results are shown in bold.

		Purity			
Data set	Method\p	0.1	0.3	0.5	0.7
BDGP	BSV	59.64±1.43	54.67±1.34	44.49±0.59	35.96±0.32
	PVC	57.87±2.68	59.01±0.04	69.10±5.04	65.83±1.68
	UEAF	90.66±0.57	89.32±0.00	87.08±0.00	76.88±0.00
	CDIMC-net	80.37±0.79	75.27±0.43	67.71±1.05	57.76±5.27
	MKKM-IK	65.01±0.03	59.80±0.00	52.56±0.00	43.84±0.00
	EE-R-IMVC	65.44±0.00	57.52±0.00	42.72±0.00	35.52±2.63
	COMPLETER	43.90±4.40	43.59±0.04	45.10±5.95	40.90±2.36
	OS-LF-IMVC	82.78±2.18	74.34±1.16	59.71±3.22	45.34±1.39
	DSIMVC	**98.40±0.26**	**96.93±0.45**	**95.29±0.37**	**92.14±0.84**
MNIST-USPS	BSV	52.77±0.11	47.66±0.14	38.91±0.11	27.82±0.11
	PVC	67.95±1.54	67.83±1.36	55.56±0.57	55.87±1.63
	UEAF	72.74±2.08	67.20±1.26	66.7±0.00	58.88±0.00
	CDIMC-net	52.25±4.52	51.00±0.23	48.63±0.51	32.44±1.78
	MKKM-IK	73.14±0.58	64.64±0.00	49.99±1.11	36.18±0.00
	EE-R-IMVC	75.12±0.48	60.38±0.00	45.64±0.00	28.02±0.00
	COMPLETER	96.87±1.04	96.59±0.82	94.40±4.14	85.49±4.86
	OS-LF-IMVC	63.61±1.18	48.62±1.98	34.95±1.50	25.52±0.12
	DSIMVC	**98.88±0.09**	**97.89±0.14**	**96.78±0.25**	**93.34±0.64**
CCV	BSV	21.76±0.25	20.06±0.28	18.52±0.19	16.79±0.16
	PVC	20.32±0.71	18.98±0.89	17.77±0.75	19.63±1.10
	UEAF	29.47±0.00	28.08±0.00	26.24±0.00	18.32±3.22
	CDIMC-net	19.10±0.79	19.96±1.11	18.05±0.66	15.79±0.95
	MKKM-IK	22.81±0.31	21.07±0.00	18.31±0.00	17.10±0.00
	EE-R-IMVC	28.43±0.19	25.28±0.00	20.12±0.00	16.36±0.00
	COMPLETER	24.02±1.10	22.46±0.61	20.82±0.87	18.67±0.75
	OS-LF-IMVC	22.99±0.79	20.18±0.48	17.61±0.47	15.55±0.36
	DSIMVC	**34.66±1.05**	**33.18±0.98**	**30.74±0.78**	**28.59±0.91**
Multi-Fashion	BSV	54.37±1.21	46.74±0.85	37.05±1.65	28.24±0.68
	PVC	47.54±3.02	52.52±1.90	48.87±0.88	51.99±0.46
	UEAF	60.71±0.86	54.33±3.17	50.48±1.26	31.12±0.00
	CDIMC-net	52.87±3.97	45.38±2.45	44.85±4.19	40.44±2.67
	MKKM-IK	70.13±0.03	59.96±0.00	47.18±0.00	30.64±0.00
	EE-R-IMVC	72.89±0.99	63.52±0.00	51.40±0.00	20.34±0.00
	COMPLETER	81.04±1.32	74.67±2.01	75.20±4.66	71.24±3.44
	OS-LF-IMVC	65.78±1.13	52.39±2.32	39.27±1.07	29.51±1.23
	DSIMVC	**88.76±0.90**	**85.02±1.73**	**83.55±2.35**	**76.55±1.85**

where $\tilde{c}_1 := m(m-1)(8D^2E^4+K+2C+2)/2$ and $\tilde{c}_2 := (8D^2E^4+K+C+1)m(m-1)/2$. Note that $\widehat{L}(f(\mathcal{D}^c;w^*))$ can be rewritten as $\widehat{L}(f(\mathcal{D}^c;w^*)) = \widehat{L}(f(\mathcal{D}^c;w(\phi')))$, as shown in the proof of Theorem 3.2. Let $\mathcal{L}(f(w^*(\hat{\phi})))$ and $\mathcal{L}(f(w^*(\phi')))$ be the expectation of $\widehat{\mathcal{L}}(f(\mathcal{D}^c;w^*(\hat{\phi})))$ and $\widehat{\mathcal{L}}(f(\mathcal{D}^c;w^*(\phi')))$, respectively. With probability at least $1-\frac{\delta}{2}$, we have

$$\mathcal{L}(f(w^*(\hat{\phi}))) \leq \widehat{\mathcal{L}}(f(\mathcal{D}^c;w^*(\hat{\phi}))) + \frac{\tilde{c}_1}{\sqrt{n}} + \tilde{c}_2\sqrt{\frac{\log 12/\delta}{n}},$$

and

$$\widehat{\mathcal{L}}(f(\mathcal{D}^c;w^*(\phi'))) \leq \mathcal{L}(f(w^*(\phi'))) + \frac{\tilde{c}_1}{\sqrt{n}} + \tilde{c}_2\sqrt{\frac{\log 12/\delta}{n}}.$$

According to Theorem 3.2, there exists a constant $\varepsilon \geq 0$ such that $\widehat{\mathcal{L}}(f(\mathcal{D}^c;w^*(\hat{\phi}))) + \varepsilon = \widehat{\mathcal{L}}(f(\mathcal{D}^c;w^*(\phi')))$ hold. Thus, with probability at least $1-\delta$, for any $f \in \mathcal{F}$, we have

$$\mathcal{L}(f(w^*(\hat{\phi}))) + \varepsilon \leq \mathcal{L}(f(w^*(\phi'))) + \frac{c_1}{\sqrt{n}} + c_2\sqrt{\frac{\log 12/\delta}{n}},$$

where $\varepsilon := \widehat{\mathcal{L}}(f(\mathcal{D}^c;w^*(\phi'))) - \widehat{\mathcal{L}}(f(\mathcal{D}^c;w^*(\hat{\phi})))$. $c_1 := 2\tilde{c}_1$ and $c_2 := 2\tilde{c}_2$ are constants dependent on D, E, K, m. This finishes the proof.

B. Experiments

In this part, we present implementation details of the proposed method. For MNIST-USPS and Multi-Fashion datasets, the raw data (*i.e.*, images) are reshaped as vectors. The hidden features are extracted by the fully connected network with the

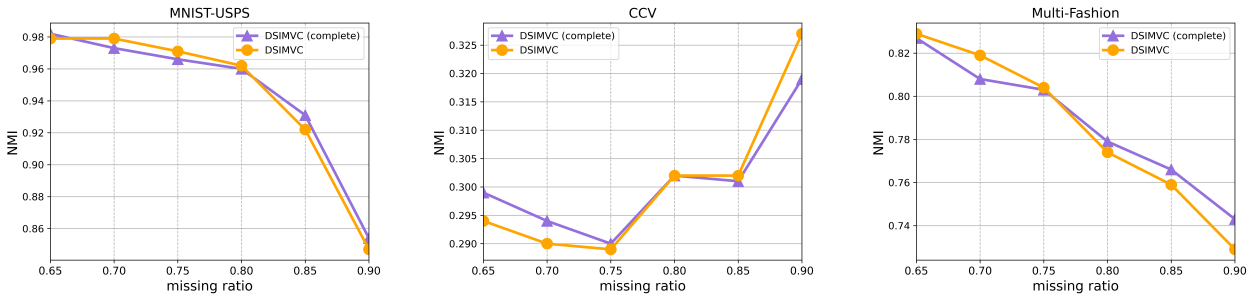


Figure 5. Clustering performance comparison in terms of NMI of DSIMVC and its variant on MNIST-USPS, CCV and Multi-Fashion.

same architecture $D^p - 500 - 500 - 2000 - 512$, where D^p is the dimensionality of the p-th view samples. Then the hidden features are fed into a two layers MLP with architecture $512 - 512 - 256$ to obtain the semantic features $f_{\mathcal{Z}}(\mathbf{x})$. Also, the cluster assignment probability $f_{\mathcal{Q}}(\mathbf{x})$ is obtained from the hidden features by another two-layers MLP with architecture $512 - 512 - K$, where K denotes the number of categories. Following (Shu et al., 2019; Guo et al., 2020), the weighting function is a one-layer MLP where the number of hidden layer's neurons is 100 and the activation function of the output layer is Sigmoid. The activation function of all hidden layers is ReLU. To accelerate the training process, mini-batch gradient descent with Adam optimizer is adopted, and the batch size is set to 256 for all datasets. As mentioned in the main paper, the trade-off parameter γ and the number of neighbors k are empirically set to 0.5 and 3, respectively. We adopt the Faiss library (Johnson et al., 2019) to search for the nearest neighbors based on learned features. The learning rate η_w and η_ϕ are set as 0.0003 and 0.0004. The parameters of the neural network are initialized by pretraining on complete data, and the neighbors are initialized by the model after pretraining. We report the results of baseline methods obtained by running the open-source code with default settings .