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

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

1. Proofs

In this part, we provide the detailed proofs of the theoretical results in the main paper.

1.1. Proof of Theorem 1

Proof. According to the definitions in the main paper, the empirical clustering risk of the model learning from data of the new increased view is denoted as

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

And the empirical clustering risk of the model learning from data before view increase is defined as

$$\hat{\mathcal{L}}_n^{\{1,\cdots,p-1\}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathcal{C}(\mathcal{F}_{\{1,\cdots,p-1\}}(\{\mathbf{x}_i^v\}_{v=1}^{p-1}))). \tag{2}$$

The objective of the optimization problem is formulated as

$$\min_{\boldsymbol{\lambda}\in\Lambda}\Big\{\min_{\boldsymbol{\theta}\in\Theta}\hat{\mathcal{L}}_n(\boldsymbol{\lambda},\boldsymbol{\theta})\Big\}, \tag{3}$$

with

$$\hat{\mathcal{L}}_n(\boldsymbol{\lambda},\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathcal{C}(\mathcal{S}(\{\mathbf{x}_i^v\}_{v=1}^{p};\{\mathcal{F}\}_p))).$$

According to the definition of $\mathcal{S}$ in the main paper, Eq. (1) and Eq. (2) can be expressed as

$$\hat{\mathcal{L}}_n^p = \hat{\mathcal{L}}_n(\boldsymbol{\lambda},\boldsymbol{\theta})|_{\lambda_1=1,\lambda_2=\lambda_3=0}, \tag{4}$$

and

$$\hat{\mathcal{L}}_n^{\{1,\cdots,p-1\}} = \hat{\mathcal{L}}_n(\boldsymbol{\lambda},\boldsymbol{\theta})|_{\lambda_2=1,\lambda_1=\lambda_3=0}. \tag{5}$$

As defined, $\hat{\mathcal{L}}_n^*$ is the optimal value of the optimization problem in Eq. (3). Thus, for all $\boldsymbol{\lambda}\in\Lambda$, the following equality holds:

$$\hat{\mathcal{L}}_n^* \le \hat{\mathcal{L}}_n(\boldsymbol{\lambda},\boldsymbol{\theta}). \tag{6}$$

Combining Eqs. (4), (5), and (6), we have

$$\hat{\mathcal{L}}_n^* \le \hat{\mathcal{L}}_n^p,$$
$$\hat{\mathcal{L}}_n^* \le \hat{\mathcal{L}}_n^{\{1,\cdots,p-1\}}.$$

Thus, the following equality holds

$$\hat{\mathcal{L}}_n^* \le \min\{\hat{\mathcal{L}}_n^p, \hat{\mathcal{L}}_n^{\{1,\cdots,p-1\}}\}.$$

This finishes the proof.

*Corresponding author.

1.2. Proof of Theorem 2

To prove Theorem 2, we first introduce the following lemma.

Lemma 1. *We define the empirical risk and its expectation as*

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

and

$$\mathcal{L}(g_{S,H}) = \frac{1}{\binom{K}{2}}\sum_{l=1}^{K-1}\sum_{s=l+1}^{K}\mathbb{E}_{\mathbf{x},\mathbf{x}'}\Big[g_{S,H_{l,s}}(\mathbf{x},\mathbf{x}')\Big] + \sum_{l=1}^{K}\mathbb{E}_{\mathbf{x},\mathbf{x}'}\Big[g_{H_l}(\mathbf{x},\mathbf{x}')\Big] + \frac{1}{\binom{K}{2}}\sum_{l=1}^{K-1}\sum_{s=l+1}^{K}\mathbb{E}_{\mathbf{x}}\Big[g_{S,H_{l,s}}(\mathbf{x},\mathbf{x})\Big].$$

For any $0<\delta<1$, with probability $1-\delta$, the following inequality holds:

$$\mathcal{L}(g_{S,H}) \leq \hat{\mathcal{L}}_n(g_{S,H}) + \frac{13KM}{\sqrt{n}} + 4(K+1)M\sqrt{\frac{\log\frac{1}{\delta}}{2n}}. \tag{7}$$

Proof. This proof is inspired by [8]. For any samples $S=\{\mathbf{x},\cdots,\mathbf{x}_n\}$, let $\bar{S}$ be a samples different from S by only one instance $\bar{\mathbf{x}}_r$. The empirical clustering risk of the hypothesis function $g_{S,H}$ on $\bar{S}$ is denoted as $\hat{\mathcal{L}}'_n$. We have

$$\begin{aligned}
&\left|\sup_{g_{S,H}\in\mathcal{G}}\left|\mathcal{L}(g_{S,H})-\hat{\mathcal{L}}_n(g_{S,H})\right| - \sup_{g_{S,H}\in\mathcal{G}}\left|\mathcal{L}(g_{S,H})-\hat{\mathcal{L}}'_n(g_{S,H})\right|\right| \\
\leq& \sup_{g_{S,H}\in\mathcal{G}}\left|\hat{\mathcal{L}}_n(g_{S,H})-\hat{\mathcal{L}}'_n(g_{S,H})\right| \\
\leq& \sup_{g_{S,H}\in\mathcal{G}}\left[\frac{2}{n^2\binom{K}{2}}\sum_{i=1,i\neq r}^{n}\left(\left|\sum_{l=1}^{K-1}\sum_{s=l+1}^{K} g_{S,H_{l,s}}(\mathbf{x}_i,\mathbf{x}_r)\right| + \left|\sum_{l=1}^{K-1}\sum_{s=l+1}^{K} g_{S,H_{l,s}}(\mathbf{x}_i,\bar{\mathbf{x}}_r)\right|\right) + \frac{1}{n^2\binom{K}{2}}\left|\sum_{l=1}^{K-1}\sum_{s=l+1}^{K} g_{S,H_{l,s}}(\mathbf{x}_r,\mathbf{x}_r)\right|\right. \\
&\left. + \frac{1}{n^2\binom{K}{2}}\left|\sum_{l=1}^{K-1}\sum_{s=l+1}^{K} g_{S,H_{l,s}}(\bar{\mathbf{x}}_r,\bar{\mathbf{x}}_r)\right| + \frac{1}{\binom{n}{2}}\sum_{i=1,i\neq r}^{n}\left(\left|\sum_{l=1}^{K} g_{H_l}(\mathbf{x}_i,\mathbf{x}_r)\right| + \left|\sum_{l=1}^{K} g_{H_l}(\mathbf{x}_i,\bar{\mathbf{x}}_r)\right|\right)\right] \\
\leq& \frac{2(2n-1)M}{n^2} + \frac{4KM}{n} \\
\leq& \frac{4(K+1)M}{n},
\end{aligned}$$

where the last inequality is obtained by the Assumption 1 in the main paper. According to the McDiarmid inequality [12], we have

$$\mathcal{L}(g_{S,H}) \leq \hat{\mathcal{L}}_n(g_{S,H}) + \mathbb{E}\sup_{g_{S,H}\in\mathcal{G}}\left|\mathcal{L}(g_{S,H})-\hat{\mathcal{L}}_n(g_{S,H})\right| + 4(K+1)M\sqrt{\frac{\log\frac{1}{\delta}}{2n}}. \tag{8}$$

Then we analyze the upper bound of the expectation term, *i.e.*, $\mathbb{E}_S\sup_{g_{S,H}\in\mathcal{G}}\left|\mathcal{L}(g_{S,H})-\hat{\mathcal{L}}_n(g_{S,H})\right|$. First, we have

$$\begin{aligned}
&\mathbb{E}_S\sup_{g_{S,h}\in\mathcal{G}}\left|\mathcal{L}(g_{S,H})-\hat{\mathcal{L}}_n(g_{S,H})\right| \\
=&\mathbb{E}_S\sup_{g_{S,H}\in\mathcal{G}}\left|\mathcal{L}(g_{S,H}) - \frac{1}{n^2\binom{K}{2}}\sum_{l=1}^{K-1}\sum_{s>l}\sum_{i=1}^{n} g_{S,H_{l,s}}(\mathbf{x}_i,\mathbf{x}_i) - \frac{1}{n^2\binom{K}{2}}\sum_{l=1}^{K-1}\sum_{s>l}\sum_{i\neq j}^{n} g_{S,H_{l,s}}(\mathbf{x}_i,\mathbf{x}_j) - \frac{1}{2\binom{n}{2}}\sum_{l=1}^{K}\sum_{i\neq j}^{n} g_{H_l}(\mathbf{x}_i,\mathbf{x}_j)\right| \\
\leq&\mathbb{E}_S\sup_{g_{S,h}\in\mathcal{G}}\left|\mathcal{L}^{(1)}(g_{S,H}) - \frac{1}{n^2\binom{K}{2}}\sum_{l=1}^{K-1}\sum_{s>l}\sum_{i=1}^{n} g_{S,H_{l,s}}(\mathbf{x}_i,\mathbf{x}_i)\right| + \mathbb{E}_S\sup_{g_{S,h}\in\mathcal{G}}\left|\mathcal{L}^{(2)}(g_{S,H}) - \frac{1}{n^2\binom{K}{2}}\sum_{l=1}^{K-1}\sum_{s>l}\sum_{i\neq j}^{n} g_{S,H_{l,s}}(\mathbf{x}_i,\mathbf{x}_j)\right| \\
&+\mathbb{E}_S\sup_{g_{S,h}\in\mathcal{G}}\left|\mathcal{L}^{(3)}(g_{S,H}) - \frac{1}{2\binom{n}{2}}\sum_{l=1}^{K}\sum_{i\neq j}^{n} g_{H_l}(\mathbf{x}_i,\mathbf{x}_j)\right|,
\end{aligned}$$

where $\mathcal{L}^{(1)}(g_{S,H}) := \frac{1}{\binom{K}{2}}\sum_{l=1}^{K-1}\sum_{s=l+1}^{K}\mathbb{E}_{\mathbf{x}}[g_{S,H_{l,s}}(\mathbf{x},\mathbf{x})]$, $\mathcal{L}^{(2)}(g_{S,H}) := \frac{1}{\binom{K}{2}}\sum_{l=1}^{K-1}\sum_{s=l+1}^{K}\mathbb{E}_{\mathbf{x},\mathbf{x}'}[g_{S,H_{l,s}}(\mathbf{x},\mathbf{x}')]$, and $\mathcal{L}^{(3)}(g_{S,H}) := \sum_{l=1}^{K}\mathbb{E}_{\mathbf{x},\mathbf{x}'}[g_{H_l}(\mathbf{x},\mathbf{x}')]$. Let $\sigma_1,\cdots,\sigma_n$ be i.i.d. Rademacher random variables taking values in $\{-1,1\}$ with

equal probability and $\bar{S} := \{\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_n\}$ be an independent copy of $S := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Then the first term can be bounded by

$$
\begin{aligned}
&\mathbb{E}_S \sup_{g_{S,h}\in\mathcal{G}} \left| \mathcal{L}^{(1)}(g_{S,H}) - \frac{1}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{n} g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_i) \right| \\
\leq &\mathbb{E}_{S,\bar{S}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{1}{n\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{n} g_{S,H_{l,s}}(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_i) - \frac{1}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{n} g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_i) \right| \\
\leq &\mathbb{E}_{S,\bar{S}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{1}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{n} [g_{S,H_{l,s}}(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_i) - g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_i)] \right| + \mathbb{E}_{\bar{S}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{n-1}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{n} g_{S,H_{l,s}}(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_i) \right| \\
= &\mathbb{E}_{S,\bar{S},\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{1}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{n} \sigma_i [g_{S,H_{l,s}}(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_i) - g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_i)] \right| + \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{n-1}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{n} \sigma_i g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_i) \right| \\
= &2\mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{1}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{n} \sigma_i g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_i) \right| + \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{n-1}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{n} \sigma_i g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_i) \right| \\
\leq &2 \max_{l,s} \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \frac{1}{n^2} \left| \sum_{i=1}^{n} \sigma_i g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_i) \right| + \max_{l,s} \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_i) \right|.
\end{aligned}
$$

The second term can be bounded by

$$
\begin{aligned}
&\mathbb{E}_S \sup_{g_{S,h}\in\mathcal{G}} \left| \mathcal{L}^{(2)}(g_{S,H}) - \frac{1}{\binom{K}{2}n^2} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i\neq j}) g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_j) \right| \\
\leq &\mathbb{E}_S \sup_{g_{S,h}\in\mathcal{G}} \left| \mathcal{L}^{(2)}(g_{S,H}) - \frac{1}{\binom{K}{2}n(n-1)} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i\neq j} g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{\binom{K}{2}n^2(n-1)} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i\neq j} g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_j) \right| \\
\leq &\mathbb{E}_S \sup_{g_{S,h}\in\mathcal{G}} \left| \mathcal{L}^{(2)}(g_{S,H}) - \frac{1}{\binom{K}{2}n(n-1)} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i\neq j} g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_j) \right| + \mathbb{E}_S \sup_{g_{S,h}\in\mathcal{G}} \left| \frac{1}{\binom{K}{2}n^2(n-1)} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i\neq j} g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_j) \right| \\
\leq &\mathbb{E}_S \sup_{g_{S,h}\in\mathcal{G}} \left| \mathcal{L}^{(2)}(g_{S,H}) - \frac{1}{\binom{K}{2}\lfloor n/2 \rfloor} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{\lfloor n/2 \rfloor} g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}) \right| + \mathbb{E}_S \sup_{g_{S,h}\in\mathcal{G}} \left| \frac{1}{\binom{K}{2}n\lfloor n/2 \rfloor} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{\lfloor n/2 \rfloor} g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}) \right| \\
\leq &\mathbb{E}_{S,\bar{S}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{1}{\binom{K}{2}\lfloor n/2 \rfloor} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{\lfloor n/2 \rfloor} [g_{S,H_{l,s}}(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_{i+\lfloor n/2 \rfloor}) - g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor})] \right| \\
&+ \mathbb{E}_S \sup_{g_{S,h}\in\mathcal{G}} \left| \frac{1}{\binom{K}{2}n\lfloor n/2 \rfloor} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{\lfloor n/2 \rfloor} g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}) \right| \\
= &\mathbb{E}_{S,\bar{S},\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{1}{\binom{K}{2}\lfloor n/2 \rfloor} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i [g_{S,H_{l,s}}(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_{i+\lfloor n/2 \rfloor}) - g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor})] \right| \\
&+ \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{1}{\binom{K}{2}n\lfloor n/2 \rfloor} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}) \right| \\
= &2\mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{1}{\binom{K}{2}\lfloor n/2 \rfloor} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}) \right| + \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{1}{\binom{K}{2}n\lfloor n/2 \rfloor} \sum_{l=1}^{K-1} \sum_{s>l} \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i g_{S,H_{l,s}}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}) \right|,
\end{aligned}
$$

where the third inequality is obtained by the Lemma A.1 in [1]. Similarly, the third term can be bounded by

$$
\begin{aligned}
&\mathbb{E}_S \sup_{g_{S,h}\in\mathcal{G}} \left| \mathcal{L}^{(3)}(g_{S,H}) - \frac{1}{n(n-1)} \sum_{l=1}^{K} \sum_{i\neq j} g_{H_l}(\mathbf{x}_i, \mathbf{x}_j) \right| \\
\leq &2\mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{l=1}^{K} \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i g_{H_l}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}) \right| \leq 2K \max_{l} \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i g_{H_l}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}) \right|.
\end{aligned}
$$

Combining the aforementioned results, according to the Khintchine-Kahane inequality [6] and the Assumption 1 in the main paper, we have

$$
\begin{aligned}
&\mathbb{E}_S \sup_{g_{S,h}\in\mathcal{G}} \left|\mathcal{L}(g_{S,H}) - \hat{\mathcal{L}}_n(g_{S,H})\right| \\
\leq& 2\max_{l,s} \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \frac{1}{n^2}\left|\sum_{i=1}^{n}\sigma_i g_{S,H_{l,s}}(\mathbf{x}_i,\mathbf{x}_i)\right| + \max_{l,s} \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i g_{S,H_{l,s}}(\mathbf{x}_i,\mathbf{x}_i)\right| \\
&+ 2\max_{l,s} \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \frac{1}{\lfloor n/2\rfloor}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\sigma_i g_{S,H_{l,s}}(\mathbf{x}_i,\mathbf{x}_{i+\lfloor n/2\rfloor})\right| + \max_{l,s} \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \frac{1}{n\lfloor n/2\rfloor}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\sigma_i g_{S,H_{l,s}}(\mathbf{x}_i,\mathbf{x}_{i+\lfloor n/2\rfloor})\right| \\
&+ 2K\max_{l} \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{g_{S,H}\in\mathcal{G}} \frac{1}{\lfloor n/2\rfloor}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\sigma_i g_{H_l}(\mathbf{x}_i,\mathbf{x}_{i+\lfloor n/2\rfloor})\right| \\
\leq& \max_{l,s} \mathbb{E}_S \sup_{g_{S,H}\in\mathcal{G}} \frac{n+2}{n^2}\left(\sum_{i=1}^{n}[g_{S,H_{l,s}}(\mathbf{x}_i,\mathbf{x}_i)]^2\right)^{\frac{1}{2}} + \max_{l,s} \mathbb{E}_S \sup_{g_{S,H}\in\mathcal{G}} \frac{2n+1}{n\lfloor n/2\rfloor}\left(\sum_{i=1}^{\lfloor n/2\rfloor}[g_{S,H_{l,s}}(\mathbf{x}_i,\mathbf{x}_{i+\lfloor n/2\rfloor})]^2\right)^{\frac{1}{2}} \\
&+ 2K\max_{l} \mathbb{E}_S \sup_{g_{S,H}\in\mathcal{G}} \frac{1}{\lfloor n/2\rfloor}\left(\sum_{i=1}^{\lfloor n/2\rfloor}[g_{H_l}(\mathbf{x}_i,\mathbf{x}_{i+\lfloor n/2\rfloor})]^2\right)^{\frac{1}{2}} \\
\leq& \frac{3M}{\sqrt{n}} + \frac{3M}{\sqrt{\lfloor n/2\rfloor}} + \frac{2KM}{\sqrt{\lfloor n/2\rfloor}} \\
\leq& \frac{13KM}{\sqrt{n}}.
\end{aligned}
$$

Incorporating this bound into Eq. (8), with probability $1-\delta$, we have

$$
\mathcal{L}(g_{S,H}) \leq \hat{\mathcal{L}}_n(g_{S,H}) + \frac{13KM}{\sqrt{n}} + 4(K+1)M\sqrt{\frac{\log\frac{1}{\delta}}{2n}}. \tag{9}
$$

Proof of Theorem 2. Without loss of generality, we assume that $\min\{\mathcal{L}^p, \mathcal{L}^{\{1,\cdots,p-1\}}\} = \mathcal{L}^p$. According to Lemma 1, with probability $1-\delta$, the following inequalities holds:

$$
\mathcal{L} - \hat{\mathcal{L}}_n^* \leq +\frac{13KM}{\sqrt{n}} + 4(K+1)M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}, \tag{10}
$$

$$
\hat{\mathcal{L}}_n^p - \mathcal{L}^p \leq +\frac{13KM}{\sqrt{n}} + 4(K+1)M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \tag{11}
$$

Combing Eq. (10) and Eq. (11), with probability $1-\delta$, the following inequality holds:

$$
\mathcal{L} + \hat{\mathcal{L}}_n^p - \hat{\mathcal{L}}_n^* \leq \mathcal{L}^p + \frac{26KM}{\sqrt{n}} + 8(K+1)M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \tag{12}
$$

According to Theorem 1, there exists a constant $\epsilon \geq 0$ such that $\hat{\mathcal{L}}_n^* + \epsilon = \min\{\hat{\mathcal{L}}_n^p, \hat{\mathcal{L}}_n^{\{1,\cdots,p-1\}}\}$ holds. Therefore, with probability $1-\delta$, we have

$$
\mathcal{L} + \epsilon \leq \min\{\mathcal{L}^p, \mathcal{L}^{\{1\cdots,p-1\}}\} + \frac{c_1}{\sqrt{n}} + c_2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}, \tag{13}
$$

where $c_1 := 26KM$ and $c_2 := 8(K+1)M$ are constants dependent on K and M. ϵ is formulated as $\epsilon := \min\{\hat{\mathcal{L}}_n^p, \hat{\mathcal{L}}_n^{\{1,\cdots,p-1\}}\} - \hat{\mathcal{L}}_n^*$. This finishes the proof.

Remark 1. *We show that when the model is trained sufficiently, the empirical clustering risk of the proposed DSMVC is an example of the divergence-based clustering framework presented in the main paper. As mentioned in the main paper, the*

empirical clustering risk of the proposed DSMVC can be rewritten as

$$\begin{aligned}\hat{\mathcal{L}}_n =& \frac{1}{\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s=l+1}^{K} \sum_{i,j=1}^{n} \frac{\mathbf{Y}_{il}\mathbf{K}_{ij}\mathbf{Y}_{js}}{\sqrt{\sum_{a,b=1}^{n} \mathbf{Y}_{al}\mathbf{K}_{ab}\mathbf{Y}_{bl} \sum_{a,b=1}^{n} \mathbf{Y}_{as}\mathbf{K}_{ab}\mathbf{Y}_{bs}}} + \frac{1}{2\binom{n}{2}} \sum_{l=1}^{K} \sum_{i,j=1,i\neq j}^{n} \mathbf{Y}_{il}\mathbf{Y}_{jl} \\ &+ \sum_{l=1}^{K-1} \sum_{s=l+1}^{K} \frac{\binom{K}{2}^{-1} \sum_{i,j=1}^{n} \mathbf{D}_{il}\mathbf{K}_{ij}\mathbf{D}_{js}}{\sqrt{\sum_{a,b=1}^{n} \mathbf{D}_{al}\mathbf{K}_{ab}\mathbf{D}_{bl} \sum_{a,b=1}^{n} \mathbf{D}_{as}\mathbf{K}_{ab}\mathbf{D}_{bs}}}.\end{aligned} \tag{14}$$

Without loss of generality, assume that $\mathbf{x}_1, \cdots, \mathbf{x}_{n_1} \in C_1; \cdots; \mathbf{x}_{n_{K-1}+1}, \cdots, \mathbf{x}_n \in C_K$, where n_1 is the number of instances that belong to the cluster C_1, n_2 is the number of instances that belong to the cluster C_2, and so on. For a given instance pairs $\{\mathbf{x}_i, \mathbf{x}_j\}$, if both the instance $\mathbf{x}_i$ and the instance $\mathbf{x}_j$ belong to the l-th cluster, then we have $\mathbf{Y}_{il}\mathbf{K}_{ij}\mathbf{Y}_{jl} \approx 1$. If instance $\mathbf{x}_i$ or instance $\mathbf{x}_j$ does not belong to the l-th cluster, then we have $\mathbf{Y}_{il} \approx 0$ or $\mathbf{Y}_{jl} \approx 0$. Thus, $\sum_{i,j=1}^{n} \mathbf{Y}_{il}\mathbf{K}_{ij}\mathbf{Y}_{jl} = n_l^2$ holds. Similarly, if both the instance $\mathbf{x}_i$ and the instance $\mathbf{x}_j$ belong to the l-th cluster, $\mathbf{D}_{il}\mathbf{K}_{ij}\mathbf{D}_{jl} \approx 1$ holds. Otherwise, we have $\mathbf{D}_{il}\mathbf{K}_{ij}\mathbf{D}_{jl} \approx 0$. Generally, the categories of the multi-view data are evenly distributed, i.e., $n_1 = \cdots = n_K = \frac{n}{K}$. Thus, the following equality group holds:

$$\begin{aligned}\frac{1}{\sqrt{\sum_{i,j=1}^{n} \mathbf{Y}_{il}\mathbf{K}_{ij}\mathbf{Y}_{jl} \sum_{i,j=1}^{n} \mathbf{Y}_{is}\mathbf{K}_{ij}\mathbf{Y}_{js}}} &= \frac{K}{n}, \\ \frac{1}{\sqrt{\sum_{i,j=1}^{n} \mathbf{D}_{il}\mathbf{K}_{ij}\mathbf{D}_{jl} \sum_{i,j=1}^{n} \mathbf{D}_{is}\mathbf{K}_{ij}\mathbf{D}_{js}}} &= \frac{K}{n}.\end{aligned} \tag{15}$$

Then Eq. (14) can be expressed as

$$\hat{\mathcal{L}}_n = \frac{K^2}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s=l+1}^{K} \sum_{i,j=1}^{n} \mathbf{Y}_{il}\mathbf{K}_{ij}\mathbf{Y}_{js} + \frac{K^2}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s=l+1}^{K} \sum_{i,j=1}^{n} \mathbf{D}_{il}\mathbf{K}_{ij}\mathbf{D}_{js} + \frac{1}{2\binom{n}{2}} \sum_{l=1}^{K} \sum_{i,j=1,i\neq j}^{n} \mathbf{Y}_{il}\mathbf{Y}_{jl}. \tag{16}$$

We define the hypothesis functions $H_{l,s}$ for $l = 1, \cdots, K-1, s = l+1, \cdots, K$, H_l for $l = 1, \cdots, K$, and S as

$$\begin{aligned}&\mathbf{Y}_{il}\mathbf{Y}_{js} + \mathbf{D}_{il}\mathbf{D}_{js} := H_{l,s}(\mathbf{x}_i, \mathbf{x}_j), \\ &\mathbf{Y}_{il}\mathbf{Y}_{jl} := H_l(\mathbf{x}_i, \mathbf{x}_j), \\ &\mathbf{K}_{ij} := S(\mathbf{x}_i, \mathbf{x}_j).\end{aligned} \tag{17}$$

Then Eq. (14) can be written as

$$\hat{\mathcal{L}}_n(g_{S,H}) = \frac{K^2}{n^2\binom{K}{2}} \sum_{l=1}^{K-1} \sum_{s=l+1}^{K} \sum_{i,j=1}^{n} S(\mathbf{x}_i, \mathbf{x}_j) H_{l,s}(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{2\binom{n}{2}} \sum_{l=1}^{K} \sum_{i,j,i\neq j}^{n} H_l(\mathbf{x}_i, \mathbf{x}_j). \tag{18}$$

Thus, neglecting a constant factor K^2 in the first term, the empirical risk of DSMVC can be regarded as an example of the divergence-based clustering framework presented in the main paper. It can be verified that $H_{l,s}(\cdot,\cdot) \in [0,2]$ for $l = 1, \ldots, K-1, s = l+1, \cdots, K$, $H_l(\cdot,\cdot) \in [0,1]$ for $l = 1, \ldots, K$, and $S(\cdot,\cdot) \in [0,1]$ hold. Thus, the defined hypothesis functions satisfy Assumption 1 in the main paper for $M = 2$. Note that training sufficiently condition presented in Lemma 1 is a sufficient rather than necessary condition. That is, the empirical clustering risk of the proposed DSMVC may still be treated as an example of this clustering framework.

2. Experiment Details

In this part, we present the experiment detail of the proposed method.

Datasets. The experiments are conducted on several benchmark multi-view datasets. **Digit** [2] consists of 2000 instances and each data point is represented by six features, including 216-D profile correlations, 76-D Fourier coefficients of the character shapes, 64-D Karhunen-Love coefficients, 6-D morphological features, 240-D pixel averages in 2×3 windows, and 47-D Zernike moments. **Caltech** [4] is consist of five features from RGB image, including 40-D wavelet moments (WM), 254-D CENTRIST, 928-D LBP, 512-D GIST, and 1984-D HOG. 200 instances are randomly sampled from each category and constructed as a multi-view dataset with 5 views. **VOC** (*PASCAL VOC 2007*) [3] contains 9,963 image-text pairs from 20 different categories. Following [16, 19], 5,649 instances are selected to construct a two-view dataset, where

Dataset	Dimensions	#Sample	#View	#Category
Caltech-2V	40, 254	1,400	2	7
Caltech-3V	40, 254, 928	1,400	3	7
Caltech-4V	40, 254, 928, 512	1,400	4	7
Caltech-5V	40, 254, 928, 512, 1,984	1,400	5	7
Digit-2V	240, 76	2,000	2	10
Digit-3V	240, 76, 216	2,000	3	10
Digit-4V	240, 76, 216, 47	2,000	4	10
Digit-5V	240, 76, 216, 47, 64	2,000	5	10
Digit-6V	240, 76, 216, 47, 64, 6	2,000	6	10
RGB-D	2,048, 300	1,449	2	13
VOC	512, 399	5,649	2	20
Multi-MNIST	28×28	60,000	2	10

Table 1. Dataset Description.

the first and the second view is 512-D Gist feature and 399-D word frequency count of the instance respectively. **RGB-D** (*SentencesNYUv2*) [5] is an indoor scenes image-text dataset where the image is described by the text. The version provided in [16, 19] is adopted in the experiments, which provides visual features from a ResNet-50 network pretrained on the ImageNet dataset and textual features from a doc2vec model pretrained on the Wikipedia dataset. **Multi-MNIST** is a multi-view version of the popular MNIST dataset [7], whose two views are the raw image and its augmented version with a highlighted edge. [16, 19]. The description of each dataset is summarized in Table 1.

Experimental settings. For a fair comparison, most settings in the experiments are the same as the settings in [16]. Concretely, the feature extractors of the proposed DSMVC on Caltech, Digit, RGB-D, and VOC datasets are implemented by fully connected layers with dimensions of 512-512-256. And the feature extractors of the proposed DSMVC on Multi-MNIST dataset are implemented by a convolution neural network, whose architecture can be expressed as Input-Conv (32, 5)-Conv (32, 5)-MaxPool (2)-Conv (32, 3)-Conv (32, 3)-MaxPool (2). Conv (32, 5) means a convolution layer where the kernel size and the number of channels are 5 and 32, respectively. MaxPool (2) denotes a maxpooling layer where the kernel size is 2. The cluster assignment module on all datasets has the same architecture, which consists of a fully connected layer with dimensions 100 and a Softmax layer. The safe module is implemented by a group of differential parameters with a Softmax activation. ReLU is adopted as the activation function. The batch size and the number of training epochs are 128 and 120 for all datasets. Adam optimizer with gradient clipping is adopted, where the max gradient norm is 5. For Digit, Caltech, RGB-D, and VOC datasets, the learning rate decay technique is adopted, where the decay step and the decay factor are 50 and 0.5, respectively. For all baseline methods, the running results of the open source codes with default settings are reported. Concretely, for spectral clustering [14], the results on the concatenation of data from all views are reported as it is a single-view clustering method. For RMVC [15], the single-view clustering results and the candidate multi-view clustering results are obtained by Best single-view normalized cut [14] and localized SimpleMKKM [11], respectively. For COMPLETER [9], results on data of the new increased view and its nearest view are reported due to the released code can only be conducted on data with two views. For other compared methods [10, 11, 13, 16–19], the settings recommended by the authors are adopted.

References

[1] Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008. 3

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

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

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

[5] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3558–3565, 2014. 6

[6] Rafał Latała and Krzysztof Oleszkiewicz. On the best constant in the khinchin-kahane inequality. *Studia Mathematica*, 109(1):101–104, 1994. 4

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

[8] Shaojie Li and Yong Liu. Sharper generalization bounds for clustering. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6392–6402, 2021. 2

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

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

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

[12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018. 2

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

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

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

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

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

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

[19] Runwu Zhou and Yi-Dong Shen. End-to-end adversarial-attention network for multi-modal clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14619–14628, 2020. 5, 6