Self-Weighted Contrastive Learning among Multiple Views for Mitigating Representation Degeneration

Jie Xu[1] **Shuo Chen**[2] **Yazhou Ren**[1] **Xiaoshuang Shi**[1]
Heng Tao Shen[1] **Gang Niu**[2] **Xiaofeng Zhu**[1,*]
[1]University of Electronic Science and Technology of China, China
[2]RIKEN Center for Advanced Intelligence Project, Japan

Abstract

Recently, numerous studies have demonstrated the effectiveness of *contrastive learning* (CL), which learns feature representations by pulling in positive samples while pushing away negative samples. Many successes of CL lie in that there exists semantic consistency between data augmentations of the same instance. In *multi-view scenarios*, however, CL might cause *representation degeneration* when the collected multiple views inherently have inconsistent semantic information or their representations subsequently do not capture sufficient discriminative information. To address this issue, we propose a novel framework called *SEM: SElf-weighted Multi-view contrastive learning with reconstruction regularization*. Specifically, SEM is a general framework where we propose to first measure the discrepancy between pairwise representations and then minimize the corresponding self-weighted contrastive loss, and thus making SEM adaptively strengthen the useful pairwise views and also weaken the unreliable pairwise views. Meanwhile, we impose a self-supervised reconstruction term to regularize the hidden features of encoders, to assist CL in accessing sufficient discriminative information of data. Experiments on public multi-view datasets verified that SEM can mitigate representation degeneration in existing CL methods and help them achieve significant performance improvements. Ablation studies also demonstrated the effectiveness of SEM with different options of weighting strategies and reconstruction terms.

1 Introduction

Contrastive learning (CL) explicitly enlarges the feature representation similarity between semantic-relevant samples, and it is adept at capturing high-level semantics while discarding irrelevant information. This learning paradigm has facilitated many research and application fields, such as visual representation [1, 2], text understanding [3, 4], and cross-modal agreement [5, 6, 7]. Samples with consistent semantics are typically constructed as positive sample pairs for CL loss (*e.g.*, InfoNCE [8]), which motivates multi-view learning scenarios [9, 10] where researchers focus on exploring common semantics among multi-view data. However, this kind of data usually is with heterogeneous views and thus cannot be directly processed by previous CL methods with two shared network branches.

To handle this situation, many *multi-view contrastive learning* (MCL) methods [11, 12, 13, 14] have been proposed, which treats multiple views as positive sample pairs and achieves important progresses in exploring multi-view common semantics (see Sec. 2 for details). Nevertheless, we find that CL might cause *representation degeneration* that the representations of high-quality views tend to degenerate. This may make the MCL methods perform worse than the optimal single view (see Sec. 3.1 and Sec. 4.1), and thus heavily limiting the usability of MCL in practical scenarios. Although

*Corresponding Author (seanzhuxf@gmail.com). Code link: *https://github.com/SubmissionsIn/SEM*.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

several CL work [15, 16] proposed different CL losses aiming at increasing robustness to noise and made important advances on vision and graph data, our experiments discover that these CL losses are still fragile in multi-view scenarios as multi-view data are with more diversity than single-view data. Different from changing CL loss, recent MCL methods [14, 17] focused on changing model structures and successfully improved the effectiveness of clustering the learned representations. Nevertheless, representation degeneration still exists in many cases and it requires further solutions.

We find that there could be two reasons leading to representation degeneration in MCL. **I)** The quality difference among multiple views. The success of CL is based on the priori condition that the constructed positive sample pair has semantic consistency, which generally holds in previous CL applications [1, 5, 8]. Unfortunately, for multi-view learning, the collected views usually have quality difference and the semantic of positive sample pairs might be inconsistent due to view diversity. Consequently, CL causes the representation degeneration of high-quality views due to the existence of low-quality views. **II)** Losing discriminative information during data processing. Multi-view data typically involve heterogeneous data forms [9, 18], *e.g.*, different dimensions, modalities, and sparsity. For achieving MCL, the model needs to transform heterogeneous multi-view data into the same form with different encoders. However, data transformation could lose discriminative information as this process has no supervised signals for maintaining information. As a result, CL might miss multiple views' common semantics and focus on semantic-irrelevant information due to inductive bias.

To this end, we propose *SElf-weighted Multi-view contrastive learning with reconstruction regularization (SEM)* as shown in Figure 1 that takes the m, n, o-th views in V views as an example (where $\mathcal{W}^{m,n}$ denotes the pairwise weight, $\mathcal{L}_{CL}^{m,n}$ is the contrastive loss, and $\mathbf{Z}^m$ is the learned representations). Specifically, SEM minimizes self-weighted contrastive losses $\mathcal{W}^{m,n}\mathcal{L}_{CL}^{m,n}$ and $\mathcal{W}^{n,o}\mathcal{L}_{CL}^{n,o}$ after measuring the discrepancy between pairwise views' representations, *i.e.*, ($\mathbf{Z}^m$,$\mathbf{Z}^n$) and ($\mathbf{Z}^n$,$\mathbf{Z}^o$), respectively. This makes SEM adaptively strengthen CL between the useful pairwise views and also weaken CL between the unreliable pairwise views. Meanwhile, SEM takes self-supervised reconstruction objectives as regularization terms ($\mathcal{R}^m$, $\mathcal{R}^n$, and $\mathcal{R}^o$) on the hidden features ($\mathbf{H}^m$, $\mathbf{H}^n$, and $\mathbf{H}^o$) of encoders for individual views, respectively. This reconstruction regularization assists CL in accessing sufficient discriminative information hidden in raw input data ($\mathbf{X}^m$, $\mathbf{X}^n$, and $\mathbf{X}^o$), which could be implemented by existing information encoder-decoder models, *e.g.*, AE [19], DAE [20], and MAE [21]. In SEM, the representations and pairwise weights are alternatively updated to mutually enhance one another.

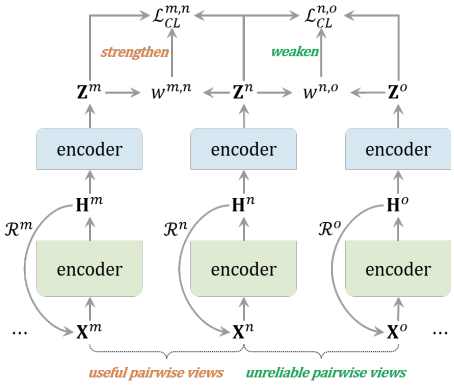
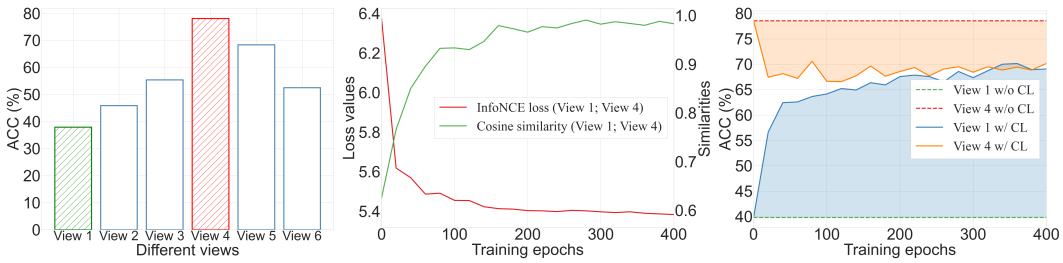


Figure 1: The framework of SEM. It leverages different networks to extract information of different views and conducts the proposed self-weighted multi-view contrastive learning with reconstruction regularization.

In summary, our contributions are: **I)** We propose a novel general framework SEM that leverages self-weighting and information reconstruction to address representation degeneration in MCL. **II)** We provide three options with different advantages to implement the weighting strategy of SEM including class mutual information, JS divergence, and maximum mean discrepancy. **III)** Theoretical and experimental analysis verified the effectiveness of SEM. It helps many CL methods (*e.g.*, InfoNCE [8], RINCE [15], and PSCL [16]) achieve significant performance improvements in multi-view scenarios.

2 Related Work

Contrastive learning (CL) As a popular self-supervised learning paradigm, CL focuses on learning semantically informative representations for downstream tasks [22, 23, 24, 25]. The most widely used loss function is InfoNCE [8] which pulls in the representations between positive sample pairs while pushing away that between negative sample pairs. Some work have attempted to explain the reasons for the success of applying InfoNCE, *e.g.*, from perspectives of mutual information [8, 26], task-dependent view [27], or deep metric learning [28, 29]. Furthermore, [30, 31] pointed out to conduct CL with reconstruction regularization to achieve robust representations for downstream tasks. RINCE [15] (a short name of Robust InfoNCE) is a variant of InfoNCE contrastive loss that considers noise in false positive sample pairs. The recent work [16] investigates CL without



(a) ACC vs. Views (b) Loss and similarity vs. Epoch (c) ACC vs. Epoch

Figure 2: (a) Clustering accuracy of individual views on Caltech dataset. (b) Contrastive loss and representation similarity between view 1 and view 4. (c) Clustering accuracy of view 1 and view 4 during contrastive learning.

conditional independence assumption on positive sample pairs and proposes a population spectral contrastive loss (we call it PSCL for short). Despite important progresses have been made, in this work, we discover that these CL losses are still fragile in multi-view scenarios where data qualities are hard to be guaranteed, and even the reconstruction regularized CL is not enough.

Multi-view contrastive learning (MCL) Different from many CL methods that usually generate two inputs by data augmentation [32], MCL aims to handle multi-view data widely exiting in real-world applications. Multi-view data often contain more than two views/modalities and they naturally form multiple inputs [33, 34, 35]. Since the semantic consistency among multiple views is not guaranteed, it is challenging to capture the useful information in multi-view data, while considering the side effects of harmful information. Therefore, MCL attracts increasing attention in recent years [36, 37, 38]. For example, CMC [11] empirically shows that MCL performed with more scene views obtains the better representations with semantic information. DCP [39] leverages the maximization of mutual information to conduct consistency learning across different views and aims to achieve a provable sufficient and minimal representation. MFLVC [14] observes the conflict between consistency and reconstruction objectives in encoder-decoder frameworks and proposes to learn multi-level features for multiple views. DSIMVC [17] establishes a theoretical framework to reduce the risk of clustering performance degradation from semantic inconsistent views. Although satisfactory results are achieved in many cases, the representation degeneration caused by CL is still not well considered and addressed. In this paper, we point out that the representation degeneration could seriously limit the application of CL in multi-view scenarios, and propose the discrepancy-based self-weighted MCL to address it.

Notations This paper leverages bold uppercase characters and bold lowercase characters to denote matrices and vectors, respectively. Operator $\|\cdot\|_2$ denotes vector ℓ_2-norm and operator $\|\cdot\|_F$ is matrix F-norm. $\{\mathbf{x}_i^v \in \mathbf{X}^v\}_{i=1,2,\ldots,N}^{v=1,2,\ldots,V}$ denotes the multi-view dataset with N samples in V views.

3 Methodology

This section first illustrates the phenomenon of representation degeneration in multi-view contrastive learning. To address this issue, we then establish a general framework of *SEM: SElf-weighed Multi-view contrastive learning with reconstruction regularization*. To implement the SEM framework, we further provide different options of weighing strategy, contrastive loss, and reconstruction term.

3.1 Motivation: Representation Degeneration in Multi-View Contrastive Learning

Researchers proposed many contrastive learning approaches and also achieved plenty of progress in multi-view learning. However, multi-view contrastive learning might result in the representation degeneration of high-quality views (*i.e.*, those views contain rich semantic information) due to the diversity of multi-view data. Specifically, we illustrate it in Figure 2 that takes a popular multi-view dataset Caltech [40] (6 views) as an example. We leverage unsupervised linear clustering accuracy obtained by K-Means [41] to evaluate the representation quality of containing class-level semantics.

Firstly, we leverage self-supervised autoencoders (the setting is shown in Appendix B) to pretrain the representations of each view's data. In Figure 2(a), one can find that different views inherently have different levels of discriminative information and exhibit different qualities, where the worst (view 1)

and the best (view 4) have a large gap. Then, we adopt InfoNCE loss to perform contrastive learning between view 1 and view 4 in Figure 2(b), and record the clustering accuracy of their representations in Figure 2(c). We can observe that InfoNCE loss is well-minimized, which makes the representation similarity (evaluated by cosine) between view 1 and view 4 converge to 1.0. The performance on view 1 gradually increases. Nevertheless, the cost is that the representations of view 4 degenerate, on which the useful discriminative information reduces and thus the performance gradually decreases.

In multi-view learning, quality difference among multiple views is a common phenomenon. However, the representation degeneration in multi-view contrastive learning might make the representations of some high-quality views tend to be mediocre and thus miss their useful discriminative information.

3.2 Self-Weighted Multi-View Contrastive Learning with Reconstruction Regularization

To mitigate representation degeneration in multi-view contrastive learning, we propose a simple but effective framework called *SEM: SElf-weighted Multi-view contrastive learning with reconstruction regularization* as shown in Figure 1. Specifically, given view-specific data $\mathbf{X}^v \in \mathbb{R}^{N\times d_v}$, we let $\mathbf{Z}^v \in \mathbb{R}^{N\times z}$ denote the corresponding new representations learned by a view-specific encoder. Between $\mathbf{X}^v$ and $\mathbf{Z}^v$, we record a precursor state of representations as $\mathbf{H}^v \in \mathbb{R}^{N\times h_v}$ (termed as hidden features), and the encoder is partitioned into two parts (the front and back parts are stacked and denoted as f^v and g^v sequentially). For the v-th view, we let Ψ^v and Φ^v denote the network parameters of f^v and g^v, respectively, and then the view-specific model can be formulated as follows:

$$\mathbf{Z}^v = g^v(\mathbf{H}^v;\Phi^v) = g^v(f^v(\mathbf{X}^v;\Psi^v);\Phi^v). \tag{1}$$

In SEM, we leverage $\mathcal{L}_{CL}^{m,n}(\mathbf{Z}^m,\mathbf{Z}^n)$ to denote a contrastive loss[2], and let $\lambda>0$ denote a trade-off coefficient on regularization terms. Then, SEM is trained by minimizing the following objective:

$$\sum\nolimits_{m,n}\mathcal{W}^{m,n}\mathcal{L}_{CL}^{m,n}(\mathbf{Z}^m,\mathbf{Z}^n) + \lambda\sum\nolimits_{v}\mathcal{R}^v(\mathbf{X}^v,\mathbf{H}^v), \tag{2}$$

where $\mathcal{W}^{m,n}$ is the pairwise weight between the m-th and the n-th views, and $\mathcal{R}^v(\mathbf{X}^v,\mathbf{H}^v)$ denotes the reconstruction regularization on $\mathbf{H}^v$. We define $\mathcal{D}(\mathbf{Z}^m,\mathbf{Z}^n)$ as the discrepancy between $\mathbf{Z}^m$ and $\mathbf{Z}^n$ and denote $\mathcal{F}$ as a negative correlation function. Then, in SEM, the pairwise weight is updated by

$$\mathcal{W}^{m,n} = \mathcal{F}(\mathcal{D}(\mathbf{Z}^m,\mathbf{Z}^n)). \tag{3}$$

Self-weighting In unsupervised settings, it is hard to know which representations within $\{\mathbf{Z}^v\}_{v=1}^V$ contain useful semantic information and which are with more noise. To mitigate the representation degeneration caused by contrastive learning, SEM needs to be adaptive to quality difference among multiple views. Therefore, different from using equal-sum manner [11, 14, 17] (*e.g.*, $\sum_{m,n}\mathcal{L}_{CL}^{m,n}$), we propose to use the pairwise weighted multi-view contrastive loss, *i.e.*, $\sum_{m,n}\mathcal{W}^{m,n}\mathcal{L}_{CL}^{m,n}$. Here, $\mathcal{W}^{m,n}$ leverages the discrepancy to achieve the adaptive self-weighting. Concretely, if two views are useful pairwise views and both with informative semantics, contrastive learning between them is adaptively strengthened; if two views are unreliable pairwise views (for example, one or two of them are with less informative semantics), contrastive learning between them is adaptively weakened.

Reconstruction regularization In Eq. (2), $\mathcal{R}^v(\mathbf{X}^v,\mathbf{H}^v)$ acts as a self-supervised objective to transfer as much discriminative information as possible from $\mathbf{X}^v$ to $\mathbf{H}^v$. When we record $\mathbf{H}^v$ as the hidden features in encoder networks, the information transfer path can be described as $\mathbf{X}^v\rightarrow\mathbf{H}^v\rightarrow\mathbf{Z}^v, v\in\{1,2,\ldots,V\}$. However, information losing might occur in the processing of $\mathbf{X}^v\rightarrow\mathbf{H}^v$ such that discriminative information from some views' data is lost, and thus making contrastive learning among $\{\mathbf{Z}^v\}_{v=1}^V$ focus on harmful noise instead of common semantics across multiple views. To this end, on hidden features $\mathbf{H}^v$, our SEM leverages $\mathbf{X}^v$ to build the reconstruction regularization $\mathcal{R}^v(\mathbf{X}^v,\mathbf{H}^v)$ to assist contrastive learning in accessing sufficient discriminative information from raw data.

[2] $\mathcal{L}_{CL}^{m,n}(\mathbf{Z}^m,\mathbf{Z}^n)$ can be easily replaced by previous contrastive losses, *e.g.*, InfoNCE [8], RINCE [15], and PSCL [16]. Let $\mathcal{P}$ denote the set of positive sample pairs and $\mathcal{N}$ is the set of negative sample pairs in the m,n-th views, q and α are hyper-parameters of RINCE, then the three contrastive losses could be formulated as follows:

$$\begin{aligned}
\mathcal{L}_{InfoNCE}^{m,n} &= -\mathbb{E}_{s^+\in\mathcal{P}}\left[s^+ - \log\left(e^{s^+}+\sum\nolimits_{s^-\in\mathcal{N}}e^{s^-}\right)\right],\\
\mathcal{L}_{RINCE}^{m,n} &= -\mathbb{E}_{s^+\in\mathcal{P}}\left[\frac{1}{q}\cdot e^{q\cdot s^+} - \frac{1}{q}\cdot\left(\alpha\cdot\left(e^{s^+}+\sum\nolimits_{s^-\in\mathcal{N}}e^{s^-}\right)\right)^q\right],\\
\mathcal{L}_{PSCL}^{m,n} &= -\mathbb{E}_{s^+\in\mathcal{P}}\left[2\cdot s^+\right] + \mathbb{E}_{s^-\in\mathcal{N}}\left[(s^-)^2\right],
\end{aligned}$$

where s^+ (s^-) denotes the cosine distance between the representations of positive (negative) sample pair.

3.3 Different Options for Implementing the SEM Framework

The crucial components of our proposed SEM as Eq. (2) include the weighting strategy $\mathcal{W}^{m,n}$, contrastive loss $\mathcal{L}_{CL}^{m,n}$, and regularization term $\mathcal{R}^v$. Next, we concentrate on the implementations of $\mathcal{W}^{m,n}$ (including JSD, MMD, and CMI) and briefly introduce the implementations of $\mathcal{L}_{CL}^{m,n}$ and $\mathcal{R}^v$.

Discrepancy measurements of weighting strategy When implementing $\mathcal{W}^{m,n} = \mathcal{F}(\mathcal{D}(\mathbf{Z}^m, \mathbf{Z}^n))$ in Eq. (3), many methods can measure the discrepancy $\mathcal{D}(\mathbf{Z}^m, \mathbf{Z}^n)$. Firstly, we can transfer representations to a probability distribution and leverage Jensen-Shannon divergence (JSD) to compute the discrepancy $\mathcal{D}_{JSD}(\mathbf{Z}^m, \mathbf{Z}^n)$. The advantages of JSD are its symmetry and simplicity, but it might be inapplicable when two distributions are non-overlapping. Furthermore, we can leverage maximum mean discrepancy (MMD) as the second method to obtain the discrepancy $\mathcal{D}_{MMD}(\mathbf{Z}^m, \mathbf{Z}^n)$. MMD can effectively measure non-overlapping two distributions, but it has higher complexity than JSD[3].

Actually, both JSD and MMD leverage all information of representations $\mathbf{Z}^m$ and $\mathbf{Z}^n$. However, the semantic-irrelevant information or random noise might also be embedded in $\mathbf{Z}^m$ and $\mathbf{Z}^n$. Moreover, what we expect to obtain is the mutual relation of their most representative semantic information. To this end, we propose Class Mutual Information (CMI) as the third method to obtain the discrepancy $\mathcal{D}_{CMI}(\mathbf{Z}^m, \mathbf{Z}^n)$. To be specific, since it is difficult to accurately estimate the mutual information (denoted as I) for multi-dimensional continuous variables $I(\mathbf{Z}^m; \mathbf{Z}^n)$, we denote $\mathbf{y}^m$ and $\mathbf{y}^n$ as 1-dimensional discrete vectors and change estimating $I(\mathbf{Z}^m; \mathbf{Z}^n)$ to computing $I(\mathbf{y}^m; \mathbf{y}^n)$ such that:

$$\mathcal{W}^{m,n} \approx \mathcal{F}(1/I(\mathbf{Z}^m; \mathbf{Z}^n)) \approx \mathcal{F}(1/I(\mathbf{y}^m; \mathbf{y}^n)), s.t. \underset{\mathbf{y}^m, \mathbf{y}^n}{\operatorname{argmax}} I(\mathbf{Z}^m; \mathbf{y}^m) + I(\mathbf{Z}^n; \mathbf{y}^n). \tag{4}$$

Intuitively, discrete class information in $\mathbf{Z}^v$ ($v \in \{m, n\}$) is 1-dimensional as well as the most representative information. Hence, we can optimize K-Means objective to extract the class information:

$$\mathbf{Y}^{v*} = \underset{\mathbf{Y}^v, \mathbf{C}^v}{\operatorname{argmax}} \|\mathbf{Z}^v - \mathbf{Y}^v \mathbf{C}^v\|_F^2, s.t. \mathbf{Y}^v (\mathbf{Y}^v)^T = \mathbf{I}_N, \mathbf{Y}^v \in \{0, 1\}^{N \times K}, \tag{5}$$

where $\mathbf{C}^v \in \mathbb{R}^{K \times z}$ denotes the K cluster centers of $\mathbf{Z}^v$. $\mathbf{Y}^{v*} \in \{0, 1\}^{N \times K}$ is the indicator matrix that can be further transformed to 1-dimensional discrete vector $\mathbf{y}^v$ by defining $y_i^v := \operatorname{argmax}_j y_{ij}^{v*}$ where $y_i^v \in \mathbf{y}^v, y_{ij}^{v*} \in \mathbf{Y}^{v*}$. In this way, the class information in $\mathbf{Z}^m$ and $\mathbf{Z}^n$ can be compressed into $\mathbf{y}^m$ and $\mathbf{y}^n$, respectively. Then, the class mutual information $I(\mathbf{y}^m; \mathbf{y}^n)$ is normalized and the discrepancy measurement $\mathcal{D}_{CMI}(\mathbf{Z}^m, \mathbf{Z}^n)$ between pairwise views is defined as follows:

$$\mathcal{D}_{CMI}(\mathbf{Z}^m, \mathbf{Z}^n) = \frac{H(\mathbf{y}^m) + H(\mathbf{y}^n)}{2 \cdot I(\mathbf{y}^m; \mathbf{y}^n)}, \tag{6}$$

where $H(\mathbf{y}^m) = -\sum_{i=1}^N p(y_i^m) \log p(y_i^m)$ is the cross-entropy of $\mathbf{y}^m$. This design of CMI has at least two advantages: 1) It is conducive to maintaining the representative class information while filtering out noise information; 2) Calculation is easy and owns better physical meaning.

Finally, it is also flexible to implement the negative correlation function $\mathcal{F}$. Considering $\mathcal{W}^{m,n} \geq 0$, we base on the three different discrepancies and simply give the following weighting strategies:

$$\begin{aligned}
\mathcal{W}_{CMI}^{m,n} &= \mathcal{F}_{CMI}(\mathcal{D}_{CMI}(\mathbf{Z}^m, \mathbf{Z}^n)) = e^{1/\mathcal{D}_{CMI}(\mathbf{Z}^m, \mathbf{Z}^n)} - 1, \\
\mathcal{W}_{JSD}^{m,n} &= \mathcal{F}_{JSD}(\mathcal{D}_{JSD}(\mathbf{Z}^m, \mathbf{Z}^n)) = e^{1 - \mathcal{D}_{JSD}(\mathbf{Z}^m, \mathbf{Z}^n)} - 1, \\
\mathcal{W}_{MMD}^{m,n} &= \mathcal{F}_{MMD}(\mathcal{D}_{MMD}(\mathbf{Z}^m, \mathbf{Z}^n)) = e^{-\mathcal{D}_{MMD}(\mathbf{Z}^m, \mathbf{Z}^n)}.
\end{aligned} \tag{7}$$

Compatibility for contrastive learning When implementing the contrastive loss $\mathcal{L}_{CL}^{m,n}$, it should be pointed out that multi-view contrastive learning usually has to handle more than two views (*i.e.*,

[3]We write $\hat{\mathbf{z}}_i^m \in \hat{\mathbf{Z}}^m = Softmax(\mathbf{Z}^m)$. $k(\mathbf{z}_i, \mathbf{z}_j)$ denotes the inner product of $\phi(\mathbf{z}_i)$ and $\phi(\mathbf{z}_j)$, where $\phi(\cdot)$ denotes the mapping (*e.g.*, by Gaussian kernel) to project representations into Reproducing Kernel Hilbert Space (RKHS). Then, $\mathcal{D}_{JSD}(\mathbf{Z}^m, \mathbf{Z}^n)$ and $\mathcal{D}_{MMD}(\mathbf{Z}^m, \mathbf{Z}^n)$ can be formulated as follows:

$$\mathcal{D}_{JSD}(\mathbf{Z}^m, \mathbf{Z}^n) = \frac{1}{2} \sum_{i=1}^N p(\hat{\mathbf{z}}_i^m) \log \left(\frac{2 \cdot p(\hat{\mathbf{z}}_i^m)}{p(\hat{\mathbf{z}}_i^m) + p(\hat{\mathbf{z}}_i^n)} \right) + \frac{1}{2} \sum_{i=1}^N p(\hat{\mathbf{z}}_i^n) \log \left(\frac{2 \cdot p(\hat{\mathbf{z}}_i^n)}{p(\hat{\mathbf{z}}_i^n) + p(\hat{\mathbf{z}}_i^m)} \right),$$

$$\mathcal{D}_{MMD}(\mathbf{Z}^m, \mathbf{Z}^n) = \frac{1}{N^2} \left[\sum_{i=1}^N \sum_{j=1}^N k(\mathbf{z}_i^m, \mathbf{z}_j^m) + \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{z}_i^n, \mathbf{z}_j^n) - 2 \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{z}_i^m, \mathbf{z}_j^n) \right].$$

$\{\mathbf{Z}^v\}_{v=1}^V, V > 2$), which is different from two-view setting (*e.g.*, $\{\mathbf{Z}^1, \mathbf{Z}^2\}$) in traditional contrastive learning. To make our SEM framework be compatible with previous contrastive learning methods, we construct positive/negative sample pairs as follows. Specifically, for two views $\{\mathbf{z}_i^m \in \mathbf{Z}^m, \mathbf{z}_j^n \in \mathbf{Z}^n\}$, the positive sample pairs are $\{\mathbf{z}_i^m, \mathbf{z}_i^n\}_{i=1,\dots,N}$; for any $\mathbf{z}_i^m$, its negative sample pairs are $\{\mathbf{z}_i^m, \mathbf{z}_j^v\}_{j\neq i}^{v=m,n}$. Cosine with a temperature parameter τ is leveraged to measure the representation distance between pairs, *i.e.*, $s = 1/\tau \cdot \langle \mathbf{z}_i^m, \mathbf{z}_j^n \rangle / \|\mathbf{z}_i^m\|_2 \|\mathbf{z}_j^n\|_2$. Then, we compute the contrastive loss between two views and sum all combinations as Eq. (2). We formulated three contrastive losses in Sec. 3.2, and the experiments in Sec. 4.1 will verify the compatibility of our SEM framework to them.

Reconstruction regularization When implementing the regularization term $\mathcal{R}^v(\mathbf{X}^v, \mathbf{H}^v)$ in Eq. (2), we are motivated by the information encoding-decoding process [14, 19, 30], and stack a view-specific decoder f_-^v with network parameter Ω^v on each view's $\mathbf{H}^v$ to perform data recovery of $\mathbf{X}^v$. In this way, the regularization term in SEM can be implemented with the reconstruction loss of autoencoders[4], whose encoder-decoder models can make hidden features preserve discriminative information of data. When decoder generally rebuilds $\mathbf{X}^v$ with $\mathbf{H}^v$, we can believe that $\mathbf{H}^v$ compresses the sufficient information of $\mathbf{X}^v$, for promoting contrastive learning fully access discriminative information of data.

Algorithm 1: Self-weighted multi-view contrastive learning with reconstruction regularization

Input: Dataset $\{\mathbf{X}^v\}_{v=1}^V$, Training epochs E, Step size S, Batch size n, Hyper-parameter λ
Initialize $\{\Psi^v, \Omega^v\}_{v=1}^V$ by Eq. (8) and initialize $\{\mathcal{W}^{m,n}\}_{m,n=1}^V$ with $\{\mathbf{H}^v\}_{v=1}^V$ like Eq. (7)
for $e \in \{1, 2, \dots, E\}$ **do**
  for $b \in \{1, 2, \dots, N/n\}$ **do**
    Pick mini-batch data $\{\{\mathbf{x}_i^v\}_{i=(b-1)n+1}^{bn}\}_{v=1}^V$ from $\{\mathbf{X}^v\}_{v=1}^V$
    Compute the gradient of loss via Eq. (2) on the mini-batch data
    Update $\{\Phi^v, \Psi^v, \Omega^v\}_{v=1}^V$ via Adam [42] optimizer
  if $\mathrm{mod}(e, S) == 0$ **then**
    Update $\{\mathcal{W}^{m,n}\}_{m,n=1}^V$ with $\{\mathbf{Z}^v\}_{v=1}^V$ by Eq. (7)
Output: Model parameters $\{\Phi^v, \Psi^v\}_{v=1}^V$

The training steps of SEM is summarized in Algorithm 1, where representations and weights are updated alternatively to make them promote each other. E denotes total training epochs, and the step size S denotes the number of training epochs after each update of pairwise weights. As we cannot obtain meaningful $\{\mathbf{y}^v\}_{v=1}^V$ before we start training neural networks, we first obtain meaningful $\{\mathbf{H}^v\}_{v=1}^V$ by pre-training the model with Eq. (8), and then initialize $\{\mathcal{W}^{m,n}\}_{m,n=1}^V$ with $\{\mathbf{H}^v\}_{v=1}^V$.

3.4 Theoretical Analysis

In this part, we theoretically analyze the mechanism of SEM in exploring mutual information among multiple views while mitigating representation degeneration. All proofs are given in Appendix A.

Considering SEM with InfoNCE loss and CMI weighting strategy, we have the following theorem indicating that minimizing the self-weighted contrastive loss keeps maximizing the mutual information between useful pairwise views, as well as avoiding the effects between unreliable pairwise views.

[4]We borrow the core ideas of information reconstruction applied in vanilla autoencoder (AE [19]), denoising autoencoder (DAE [20]), and masked autoencoder (MAE [21]) and provide three reconstruction regularization options. In a same form, the three kinds of reconstruction loss functions could be formulated as follows:

$$\begin{aligned}
\mathcal{R}_{AE}^v(\mathbf{X}^v, \mathbf{H}^v) &= \|\mathbf{X}^v - f_-^v(\mathbf{H}^v; \Omega^v)\|_F^2 = \|\mathbf{X}^v - f_-^v(f^v(\mathbf{X}^v; \Psi^v); \Omega^v)\|_F^2, \\
\mathcal{R}_{DAE}^v(\mathbf{X}^v, \tilde{\mathbf{H}}^v) &= \left\|\mathbf{X}^v - f_-^v(\tilde{\mathbf{H}}^v; \Omega^v)\right\|_F^2 = \|\mathbf{X}^v - f_-^v(f^v(\mathbf{X}^v + \epsilon; \Psi^v); \Omega^v)\|_F^2, \\
\mathcal{R}_{MAE}^v(\mathbf{X}^v, \ddot{\mathbf{H}}^v) &= \left\|\mathbf{X}^v - f_-^v(\ddot{\mathbf{H}}^v; \Omega^v)\right\|_F^2 = \|\mathbf{X}^v - f_-^v(f^v(\mathbf{X}^v \odot \mathbf{A}; \Psi^v); \Omega^v)\|_F^2,
\end{aligned} \tag{8}$$

where $\mathbf{X}^v + \epsilon$ denotes the data disturbed by random Gaussian noise $\epsilon \in \mathbb{R}^{N \times d_v}$ in DAE. $\mathbf{X}^v \odot \mathbf{A}$ is the data masked by random $0-1$ matrix $\mathbf{A} \in \{0, 1\}^{N \times d_v}$ in MAE. $\tilde{\mathbf{H}}^v$ and $\ddot{\mathbf{H}}^v$ denote the representations inferred from data $\mathbf{X}^v + \epsilon$ and $\mathbf{X}^v \odot \mathbf{A}$ in DAE and MAE, respectively.

Theorem 1. *For any three views ($v \in \{m, n, o\}$), if class mutual information only exists in two views, e.g., $I(\mathbf{y}^m; \mathbf{y}^o) \to 0$, $I(\mathbf{y}^n; \mathbf{y}^o) \to 0$, and $I(\mathbf{y}^m; \mathbf{y}^n) = \delta$, $\delta > 0$, we have minimizing the weighted InfoNCE losses $\mathcal{W}^{m,n}\mathcal{L}^{m,n}_{InfoNCE}(\mathbf{Z}^m, \mathbf{Z}^n) + \mathcal{W}^{m,o}\mathcal{L}^{m,o}_{InfoNCE}(\mathbf{Z}^m, \mathbf{Z}^o) + \mathcal{W}^{n,o}\mathcal{L}^{n,o}_{InfoNCE}(\mathbf{Z}^n, \mathbf{Z}^o)$ is equivalent to maximizing the mutual information between the two views $(e^{\delta/\log N} - 1)I(\mathbf{Z}^m; \mathbf{Z}^n)$.*

Combining with the information losing of each layer through encoder networks, the following theorem further reveals that reconstruction regularization on the hidden features $\mathbf{H}^v$ is conducive to alleviating the losing of discriminative semantic information through data transformation. Hence, we treat the layer output closest to $\mathbf{Z}^v$ in encoders as hidden features to maximize $\prod_{l=t^m+1}^{L^m}(1-\gamma_l^m)$ and $\prod_{l=t^n+1}^{L^n}(1-\gamma_l^n)$, aiming at maintaining useful semantic information for contrastive learning.

Theorem 2. *For any two views ($v \in \{m, n\}$) with positive class mutual information, denoting L^v as the total layer number of the v-th view's encoder network before representation $\mathbf{Z}^v$, the l-th layer has the information losing rate $\gamma_l^v \geq 0$. If $\mathbf{S}$ is an oracle variable that contains and only contains multiple views' discriminative semantic information, and $\mathbf{H}^v$ is the t^v-th layer's features, we have minimizing the regularized loss $\mathcal{W}^{m,n}\mathcal{L}^{m,n}_{InfoNCE}(\mathbf{Z}^m, \mathbf{Z}^n) + \lambda \sum_v \mathcal{R}^v(\mathbf{X}^v, \mathbf{H}^v)$ is expected to obtain $I(\mathbf{S}; \mathbf{Z}^m; \mathbf{Z}^n) \leq \min\{I(\mathbf{S}; \mathbf{X}^m) \cdot \prod_{l=t^m+1}^{L^m}(1-\gamma_l^m), I(\mathbf{S}; \mathbf{X}^n) \cdot \prod_{l=t^n+1}^{L^n}(1-\gamma_l^n)\}$.*

4 Experiments

This section validates the effectiveness of our SEM. Specifically, we first conduct comparison experiments on state-of-the-art contrastive learning baselines and SEM with three options of contrastive losses (*i.e.*, $\mathcal{L}_{InfoNCE}, \mathcal{L}_{PSCL}, \mathcal{L}_{RINCE}$). We then conduct ablation studies with three options of weighting strategies (*i.e.*, $\mathcal{W}_{CMI}, \mathcal{W}_{JSD}, \mathcal{W}_{MMD}$), as well as with three options of reconstruction terms (*i.e.*, $\mathcal{R}_{AE}, \mathcal{R}_{DAE}, \mathcal{R}_{MAE}$). Evaluation is built on the concatenation of all views' representations learned by methods. Finally, we show SEM's training process and its hyper-parameter analysis. We provided more experimental results as well as all implementation details of SEM in Appendix.

Datasets Our experiments employ five open-source multi-view datasets. Their information is shown in Table 1, where DHA [43] is a depth-included human action dataset where each action has RGB and depth features; CCV [44] refers to the columbia consumer video database whose samples are described with SIFT, STIP, and MFCC features; NUSWIDE [45] collects web images with multiple views (color histogram, block-wise color moments, color correlogram, edge direction histogram, and wavelet texture); Caltech [40] is a widely-used image dataset which leverages six views (Gabor, Wavelet moments, CENTRIST, HOG, GIST, and LBP) to represent samples; YoutubeVideo [46] is a large-scale dataset where each sample has three views including cuboids histogram, HOG, and vision misc. These datasets are diverse in forms and are often organized to comprehensively evaluate the performance of multi-view methods.

Table 1: Information of datasets

Name	View	Size	Class
DHA	2	483	23
CCV	3	6,773	20
NUSWIDE	5	5,000	5
Caltech	6	1,400	7
YoutubeVideo	3	101,499	31

4.1 Comparison Experiments on Contrastive Learning

Baselines K-Means-BSV denotes K-Means clustering results on the best single-view of raw data, and we leverage this baseline to investigate the representation degeneration in comparison methods. InfoNCE [8], PSCL [16], and RINCE [15] are three kinds of CL methods. Since their original versions are designed to handle single views, we extended them to multi-view scenarios as did in [11, 17]. CMC [11], DCP [39], MFLVC [14], and DSIMVC [17] are four kinds of MCL methods. We evaluate our SEM with different contrastive losses (*i.e.*, SEM+InfoNCE, SEM+PSCL, and SEM+RINCE), where the weighting strategy and reconstruction term are fixed to $\mathcal{W}_{CMI}$ and $\mathcal{R}_{AE}$, respectively.

We leverage the linear clustering method K-Means to evaluate the performance of learning representations and report the average results of 10 runs in Table 2. The results indicate that: **I)** Our SEM framework is compatible with different contrastive losses (*e.g.*, InfoNCE, PSCL, and RINCE) and we can clearly observe that SEM+InfoNCE/PSCL/RINCE successfully improve the baselines for large margins. For instance, SEM+InfoNCE respectively outperforms InfoNCE by about 25%, 13%, 4%, 7%, 12% ACC on the five datasets. **II)** MCL approaches could access the semantic information from multiple views, and thus outperforming that from single views. However, a side effect is

Table 2: Linear clustering performance evaluated by ACC and NMI (mean±std%)

Method	DHA		CCV		NUSWIDE		Caltech		YoutubeVideo	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
K-Means-BSV	66.6±2.6	78.0±1.3	19.5±0.3	17.8±0.3	39.7±0.0	11.6±0.0	85.1±0.1	75.6±0.1	16.5±0.6	15.9±0.4
InfoNCE [8]	54.9±3.8	77.7±2.1	25.8±0.9	25.9±0.7	56.3±2.2	30.3±1.4	79.9±1.0	71.4±0.9	19.6±0.1	19.7±0.0
PSCL [16]	39.8±2.8	72.9±2.0	21.5±1.6	24.3±1.0	53.4±0.4	27.4±0.5	67.8±2.8	69.5±2.8	15.4±0.3	14.3±0.5
RINCE [15]	49.9±6.2	76.3±2.6	22.5±0.4	23.5±0.2	56.6±1.4	30.8±1.3	80.3±2.2	72.0±2.2	14.7±0.3	13.6±0.2
CMC [11]	65.0±2.1	79.2±1.3	21.3±0.4	21.8±0.6	56.2±1.5	24.7±0.9	72.7±1.4	60.3±1.4	19.4±0.3	19.6±0.1
DCP [39]	69.8±2.2	82.9±1.6	24.1±1.2	20.6±0.9	48.1±1.4	24.5±1.1	69.6±6.6	66.2±5.3	14.0±0.3	12.3±0.4
MFLVC [14]	70.7±1.4	81.4±0.8	31.6±0.0	31.3±0.0	55.9±0.0	27.4±0.0	77.1±0.5	67.1±0.6	18.3±0.1	18.7±0.2
DSIMVC [17]	63.8±3.0	77.2±1.7	31.8±0.9	30.8±0.6	56.7±2.3	28.0±1.6	76.9±1.7	67.3±1.3	18.9±0.3	18.7±0.2
SEM+InfoNCE	**80.9**±1.9	**84.1**±0.9	**39.4**±0.7	35.5±0.4	60.4±0.4	34.9±0.7	**87.2**±0.3	**80.3**±0.5	31.3±1.1	31.1±0.9
SEM+PSCL	69.7±4.2	81.4±1.6	39.3±1.1	**35.9**±0.6	57.8±1.4	32.6±0.9	86.3±1.7	78.6±2.1	**32.2**±0.7	**32.2**±0.6
SEM+RINCE	76.3±1.3	82.8±0.7	38.9±0.9	34.6±0.5	**60.6**±0.7	**35.6**±1.3	85.4±1.4	76.7±2.2	29.8±0.5	29.5±0.5

that contrastive learning directly increases the feature representation similarity of multiple views, which might obscure useful discriminative information hidden in high-quality views and lead to the representation degeneration. For example, on DHA and Caltech, results on many MCL methods (*e.g.*, PSCL, RINCE, CMC, and MFLVC) are worse than the single-view baseline K-Means-BSV. **III)** Our SEM not only outperforms all these MCL methods but also mitigates the representation degeneration in MCL, *e.g.*, SEM+PSCL respectively outperforms K-Means-BSV by about 3%, 20%, 18%, 1%, 16% ACC on the five datasets. This is because the framework of SEM is adaptive to multiple views' qualities, which can reduce the side effect between unreliable views with inconsistent information, for better extracting discriminative information and consistent semantics among useful views.

Additionally, we leverage the linear classification method SVM [47] to evaluate the performance of MCL methods to learn representations, where we only use 30% of the learned representations for the training set and the rest for the test set. Figures 3 and 4 show the classification performance on DHA and CCV, respectively. Our SEM improves the baseline methods (especially for PSCL and RINCE) and consistently outperforms other MCL methods (such as CMC, DCP, MFLVC, and DSIMVC). Since contrastive learning usually discards the information which is irrelevant to optimization objectives, the results further indicate that the representations learned by MCL are classification-friendly, which generally focus on catching class-level semantics among multiple views.

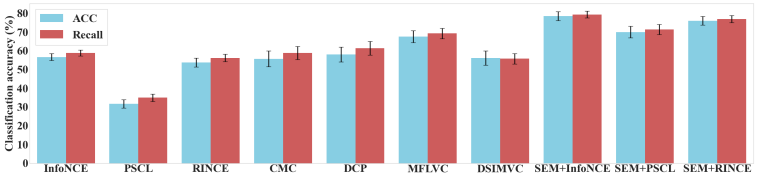


Figure 3: Classification performance on DHA.

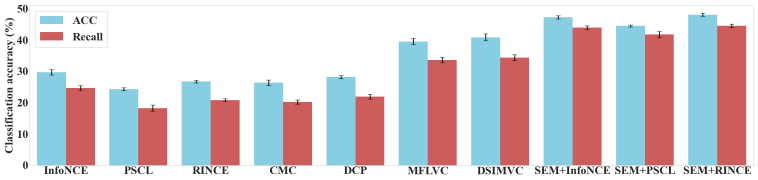


Figure 4: Classification performance on CCV.

4.2 Ablation Experiments on Self-Weighting Strategy and Reconstruction Regularization

This part presents the ablation experiments to investigate the effectiveness of different weighting strategies $\mathcal{W}_{CMI/JSD/MMD}$ and reconstruction terms $\mathcal{R}_{AE/DAE/MAE}$ in our SEM framework.

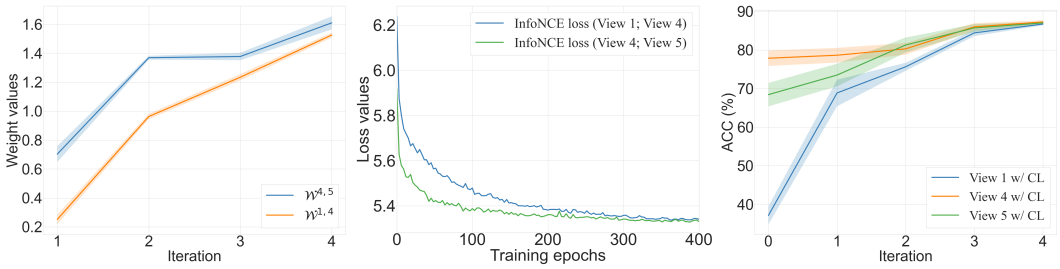
Table 3: Clustering accuracy (%) of SEM with different options of weighting strategy $\mathcal{W}$ on two datasets

	DHA	CCV
SEM w/o $\mathcal{W}$	71.3	33.5
SEM w/ $\mathcal{W}_{CMI}$	80.9 (↑ 9.6)	39.4 (↑ 5.9)
SEM w/ $\mathcal{W}_{JSD}$	80.5 (↑ 9.2)	35.6 (↑ 2.1)
SEM w/ $\mathcal{W}_{MMD}$	84.4 (↑ 13.1)	33.9 (↑ 0.4)

Table 4: Clustering accuracy (%) of SEM with different options of reconstruction term $\mathcal{R}$ on two datasets

	DHA	CCV
SEM w/o $\mathcal{R}$	60.5	28.7
SEM w/ $\mathcal{R}_{AE}$	80.9 (↑ 20.4)	39.4 (↑ 10.7)
SEM w/ $\mathcal{R}_{DAE}$	81.5 (↑ 21.0)	38.4 (↑ 9.7)
SEM w/ $\mathcal{R}_{MAE}$	83.0 (↑ 22.5)	39.5 (↑ 10.8)

Table 3 reports the linear clustering performance (evaluated by ACC) of our SEM framework without self-weighting strategy (*i.e.*, SEM w/o $\mathcal{W}$) and that with three weighting strategies (*i.e.*,



(a) Weight $vs.$ Iteration (b) Loss $vs.$ Epoch (c) ACC $vs.$ Iteration

Figure 5: (a) The change trend of weights $\mathcal{W}^{1,4}$ and $\mathcal{W}^{4,5}$ in SEM. (b) Loss values $\mathcal{L}^{1,4}_{InfoNCE}$ and $\mathcal{L}^{4,5}_{InfoNCE}$ during contrastive learning. (c) Clustering accuracy on the learned representations of view 1, view 4, and view 5.

$\mathcal{W}_{CMI}, \mathcal{W}_{JSD}, \mathcal{W}_{MMD}$), where the contrastive loss and reconstruction term are fixed to $\mathcal{L}_{InfoNCE}$ and $\mathcal{R}_{AE}$, respectively. Compared with SEM w/o $\mathcal{W}$ (this setting is the reconstruction regularized multi-view contrastive learning) that equally treats contrastive learning between any two views, SEM w/ $\mathcal{W}_{CMI/JSD/MMD}$ can adaptively weight the contrastive learning according to specific two views and thus all these three variants of SEM obtain significant improvements. For example, SEM w/ $\mathcal{W}_{MMD}$ has a 13.1% improvement on DHA and SEM w/ $\mathcal{W}_{CMI}$ has a 5.9% improvement on CCV. Results on more datasets and time costs are shown in Appendix C, where we find that the proposed weighting strategy of class mutual information $\mathcal{W}_{CMI}$ generally achieves the best performance on accuracy and time consumption among the three options of weighting strategy.

Table 4 reports the linear clustering performance (evaluated by ACC) of our SEM framework without reconstruction regularization (*i.e.*, SEM w/o $\mathcal{R}$) and that with three reconstruction terms (*i.e.*, $\mathcal{R}_{AE}, \mathcal{R}_{DAE}, \mathcal{R}_{MAE}$), where the contrastive loss and weighting strategy are fixed to $\mathcal{L}_{InfoNCE}$ and $\mathcal{W}_{CMI}$, respectively. We can easily find that the proposed SEM with reconstruction terms obviously outperforms that without reconstruction terms. For instance, compared with SEM w/o $\mathcal{R}$, SEM w/ $\mathcal{W}_{MAE}$ has 22.5% and 10.8% improvements on DHA and CCV, respectively. This is because the reconstruction regularization makes the hidden features $\{\mathbf{H}^v\}_{v=1}^V$ avoid losing discriminative information, which promotes the multi-view contrastive learning performed on subsequent $\{\mathbf{Z}^v\}_{v=1}^V$. Meanwhile, SEM w/ $\mathcal{R}_{MAE}$ and SEM w/ $\mathcal{R}_{DAE}$ perform better than SEM w/ $\mathcal{R}_{AE}$. This is because, compared with vanilla AE, DAE or MAE (by adding noise or masking on raw data) can make our model more conducive to removing semantic-irrelevant noise as well as capturing hidden patterns.

4.3 Experimental Analysis on Mechanism of SEM

This part presents the visualization and analysis on SEM to give an intuition of its behavior and mechanism, where the combination of $\mathcal{L}_{InfoNCE}$+$\mathcal{W}_{CMI}$+$\mathcal{R}_{AE}$ is taken as an example.

Let's first recall the views of Caltech dataset in Figure 2(a), we can consider that view 4 and view 5 are high-quality views, while view 1 is a low-quality view. The performance relation among them is $\text{ACC}_{\text{view 4}} > \text{ACC}_{\text{view 5}} > \text{ACC}_{\text{view 1}}$. In Figure 2(c), view 4's representation degeneration occurs.

Figure 5 shows the pairwise weights, losses, and clustering accuracy on Caltech dataset during SEM's training process, where 1 iteration corresponds to 100 epochs, *i.e.*, the step size is set to 100 epochs. Our SEM is a self-weighted multi-view contrastive learning framework that automatically infers different weights for different pairwise views as shown in Figure 5(a), where we can observe that weights $\mathcal{W}^{4,5} > \mathcal{W}^{1,4}$ and they were dynamically updated for 4 times. As a result, contrastive learning between view 4 and view 5 is strengthened by $\mathcal{W}^{4,5}$, while contrastive learning between view 1 and view 4 is weakened by $\mathcal{W}^{1,4}$. Meanwhile, loss $\mathcal{L}^{4,5}_{InfoNCE}$ is minimized earlier than loss $\mathcal{L}^{1,4}_{InfoNCE}$ as shown in Figure 5(b). In other words, since the mutual effect between view 4 and view 5 is strengthened, the effect of view 1 on view 4/view 5 is weakened such that view 4/view 5 does not degenerate. At the same time, the effect of view 4/view 5 on view 1 remains and promotes the representation learning of view 1. Consequently, all views' performance in Figure 5(c) increases through our SEM, and the representation degeneration of view 4 occurring in Figure 2(c) is mitigated.

Hyper-parameter analysis Since different datasets have different levels of reconstruction errors, the trade-off coefficient λ is introduced to balance the contrastive learning and information recovery in our SEM framework. In Figure 6(a), we change λ within the range of $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]$ and report the clustering accuracy tested on representations. The experimental results indicate that SEM is not sensitive to λ in $[10^{-1}, 10^{1}]$. In our experiments, λ is consistently set to 1 for all the five datasets. Regarding self-supervised learning, frameworks with fewer manually set hyper-parameters might be more convenient for their applications.

Additionally, we investigate the effect of cluster number when the weight strategy of our SEM framework is selected as $\mathcal{W}_{CMI}$ which needs to pre-define the cluster number when applying K-Means algorithm. As shown in Figure 6(b), when computing the class mutual information, we change the number of clusters within the range of $[K/2, K, 2K, 4K]$ where K denotes the truth class number of multi-view datasets. Compared with K, $K/2$ leads to more coarse-grained class mutual information, while $2K$ and $4K$ come in more fine-grained class mutual information. The experimental results demonstrate that SEM with $\mathcal{W}_{CMI}$ is not sensitive to the choices of cluster number.

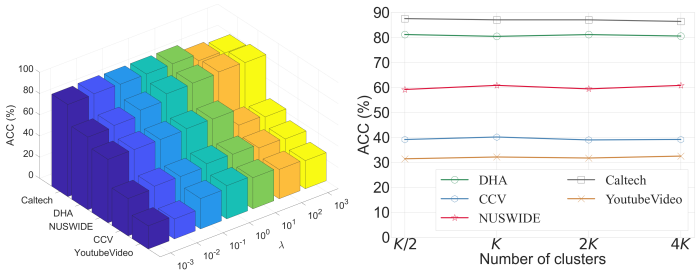


Figure 6: (a) ACC $vs.$ λ. (b) ACC $vs.$ K.

5 Conclusion

In this paper, we showcase that the representation degeneration could seriously limit the application of contrastive learning in multi-view scenarios. To mitigate this issue, we propose self-weighted multi-view contrastive learning with reconstruction regularization (SEM), which is a general framework that is compatible with different options of the contrastive loss, weighting strategy, and reconstruction term. Theoretical and experimental analysis verified the effectiveness of SEM, and it can significantly improve many existing contrastive learning methods in multi-view scenarios. Moreover, ablation studies indicated that SEM is effective with different weighting strategies and reconstruction terms.

Our future work is to extend the proposed SEM to be useful not only for multi-view scenarios, but also for other contrastive learning based domains, such as contrastive learning in sequences. Conceptually, the limitation of the self-weighting strategy is that it is more effective when there are over two views. When there are only two views, the self-weighted multi-view contrastive learning framework transforms into traditional contrastive learning but with reconstruction regularization. Therefore, another future work is to extend the view-level weighting of SEM to sample-level weighting.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFA1004100, in part by the Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China under Grant ZYGX2022YGRH009 and Grant ZYGX2022YGRH014, in part by the National Natural Science Foundation of China under Grant 62276052.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020.

[2] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[3] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, 2021.

[4] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[6] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021.

[7] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004, 2021.

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

[9] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[10] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206, 2021.

[11] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, pages 776–794, 2020.

[12] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pages 4116–4126, 2020.

[13] Erlin Pan and Zhao Kang. Multi-view contrastive graph clustering. *Advances in Neural Information Processing Systems*, 34:2148–2159, 2021.

[14] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16051–16060, 2022.

[15] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16670–16681, 2022.

[16] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

[17] Huayi Tang and Yong Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International Conference on Machine Learning*, pages 21090–21110, 2022.

[18] Yang Yang, Chubing Zhang, Yi-Chu Xu, Dianhai Yu, De-Chuan Zhan, and Jian Yang. Rethinking label-wise cross-modal retrieval from a semantic sharing perspective. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3300–3306, 2021.

[19] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in Neural Information Processing Systems*, 6:3–10, 1993.

[20] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pages 1096–1103, 2008.

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[22] Yunfan Li, Mouxing Yang, Dezhong Peng, Taihao Li, Jiantao Huang, and Xi Peng. Twin contrastive learning for online clustering. *International Journal of Computer Vision*, 130(9):2205–2221, 2022.

[23] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in Neural Information Processing Systems*, 33:8765–8775, 2020.

[24] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33:5812–5823, 2020.

[25] Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3877–3889, 2022.

[26] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

[27] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

[28] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems*, 29:1857–1865, 2016.

[29] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.

[30] Shuo Chen, Chen Gong, Jun Li, Jian Yang, Gang Niu, and Masashi Sugiyama. Learning contrastive embedding in low-dimensional space. *Advances in Neural Information Processing Systems*, 35:6345–6357, 2022.

[31] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16041–16050, 2022.

[32] A Jaiswal, A Ramesh Babu, M Zaki Zadeh, D Banerjee, and F Makedon. A survey on contrastive self-supervised learning. *Machine Learning*, 12:4182–4192, 2020.

[33] Changqing Zhang, Zongbo Han, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu, et al. CPM-Nets: Cross partial multi-view networks. *Advances in Neural Information Processing Systems*, 32:559–569, 2019.

[34] Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*, pages 17782–17800, 2022.

[35] Yang Yang, Jingshuai Zhang, Fan Gao, Xiaoru Gao, and Hengshu Zhu. DOMFN: A divergence-orientated multi-modal fusion network for resume assessment. In *Proceedings of the ACM International Conference on Multimedia*, pages 1612–1620, 2022.

[36] Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu, and Xi Peng. Partially view-aligned representation learning with noise-robust contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1134–1143, 2021.

[37] Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas Funkhouser, and Li Yi. Contrastive multimodal fusion with tupleinfonce. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 754–763, 2021.

[38] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Qing Liao, and Yuanqing Xia. Contrastive multi-view kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9552–9566, 2023.

[39] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2022.

[40] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004.

[41] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[42] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[43] Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yung-Huan Hsieh, and Hong-Ming Chen. Human action recognition and retrieval using sole depth information. In *Proceedings of the ACM International Conference on Multimedia*, pages 1053–1056, 2012.

[44] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 1–8, 2011.

[45] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 1–9, 2009.

[46] Omid Madani, Manfred Georg, and David A. Ross. On using nearly-independent feature families for high precision and confidence. *Machine Learning*, 92:457–477, 2013.

[47] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

Appendix for Self-Weighted Contrastive Learning among Multiple Views for Mitigating Representation Degeneration

Jie Xu[1] **Shuo Chen**[2] **Yazhou Ren**[1] **Xiaoshuang Shi**[1]
Heng Tao Shen[1] **Gang Niu**[2] **Xiaofeng Zhu**[1,*]
[1]University of Electronic Science and Technology of China, China
[2]RIKEN Center for Advanced Intelligence Project, Japan
[*]Corresponding Author (seanzhuxf@gmail.com)

We provide supplementary materials for the submission of *Self-Weighted Contrastive Learning among Multiple Views for Mitigating Representation Degeneration*. Specifically, Appendix A (Page-1) shows all theoretical proofs and complexity analysis of SEM; Appendix B (Page-7) includes the settings in experiments; Appendix C (Page-8) lists additional experimental results and provides more experimental analysis, which are not shown in the paper due to space; Appendix D (Page-10) discusses the limitations and future work of this paper. The code implementation, trained models, and datasets used in our method are provided in *https://github.com/SubmissionsIn/SEM*.

Appendix A Theoretical Analysis

Theorem 1. *For any three views ($v \in \{m, n, o\}$), if class mutual information only exists in two views, e.g., $I(\mathbf{y}^m; \mathbf{y}^o) \to 0$, $I(\mathbf{y}^n; \mathbf{y}^o) \to 0$, and $I(\mathbf{y}^m; \mathbf{y}^n) = \delta$, $\delta > 0$, we have minimizing the weighted InfoNCE losses $\mathcal{W}^{m,n}\mathcal{L}^{m,n}_{InfoNCE}(\mathbf{Z}^m, \mathbf{Z}^n)+\mathcal{W}^{m,o}\mathcal{L}^{m,o}_{InfoNCE}(\mathbf{Z}^m, \mathbf{Z}^o)+\mathcal{W}^{n,o}\mathcal{L}^{n,o}_{InfoNCE}(\mathbf{Z}^n, \mathbf{Z}^o)$ is equivalent to maximizing the mutual information between the two views $(e^{\delta/\log N} - 1)I(\mathbf{Z}^m; \mathbf{Z}^n)$.*

Proof. According to Proposition 1, minimizing the weighted InfoNCE losses becomes maximizing the following weighted mutual information:

$$\mathcal{W}^{m,n}I(\mathbf{Z}^m; \mathbf{Z}^n) + \mathcal{W}^{m,o}I(\mathbf{Z}^m; \mathbf{Z}^o) + \mathcal{W}^{n,o}I(\mathbf{Z}^n; \mathbf{Z}^o). \tag{1}$$

Furthermore, based on the definition of CMI weighing strategy, we have

$$\mathcal{W}^{m,n} = e^{\frac{2\cdot I(\mathbf{y}^m;\mathbf{y}^n)}{H(\mathbf{y}^m)+H(\mathbf{y}^n)}} - 1. \tag{2}$$

If $I(\mathbf{y}^m; \mathbf{y}^o) \to 0$ and $I(\mathbf{y}^n; \mathbf{y}^o) \to 0$, we obtain

$$\begin{aligned} &\lim_{I(\mathbf{y}^m;\mathbf{y}^o)\to 0} \mathcal{W}^{m,o}I(\mathbf{Z}^m; \mathbf{Z}^o) + \lim_{I(\mathbf{y}^n;\mathbf{y}^o)\to 0} \mathcal{W}^{n,o}I(\mathbf{Z}^n; \mathbf{Z}^o) \\ &= \lim_{\mathcal{W}^{m,o}\to 0} \mathcal{W}^{m,o} \cdot I(\mathbf{Z}^m; \mathbf{Z}^o) + \lim_{\mathcal{W}^{n,o}\to 0} \mathcal{W}^{n,o} \cdot I(\mathbf{Z}^n; \mathbf{Z}^o) = 0. \end{aligned} \tag{3}$$

Then, if $I(\mathbf{y}^m; \mathbf{y}^n) = \delta$, $\delta \in \mathbb{R}^+$, Eq. (1) becomes

$$\left(e^{\frac{2\cdot\delta}{H(\mathbf{y}^m)+H(\mathbf{y}^n)}} - 1\right) \cdot I(\mathbf{Z}^m; \mathbf{Z}^n). \tag{4}$$

For $H(\mathbf{y}^m) + H(\mathbf{y}^n)$, it has a maximum value $2\log N$ if $\mathbf{y}^m$ and $\mathbf{y}^n$ follow the uniform distribution, *i.e.*, $H(\mathbf{y}^m) + H(\mathbf{y}^n) = -\sum_{i=1}^N p(y_i^m)\log p(y_i^m) - \sum_{i=1}^N p(y_i^n)\log p(y_i^n) = -\sum_{i=1}^N p(y_i^m)\log\frac{1}{|\mathbf{y}^m|} - \sum_{i=1}^N p(y_i^n)\log\frac{1}{|\mathbf{y}^n|} = 2\log N$. Therefore, we have

$$\left(e^{\frac{2\cdot\delta}{H(\mathbf{y}^m)+H(\mathbf{y}^n)}} - 1\right) \cdot I(\mathbf{Z}^m; \mathbf{Z}^n) \geq \left(e^{\delta/\log N} - 1\right) \cdot I(\mathbf{Z}^m; \mathbf{Z}^n), \tag{5}$$

which completes the proof. □

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Theorem 2. *For any two views ($v \in \{m, n\}$) with positive class mutual information, denoting L^v as the total layer number of the v-th view's encoder network before representation $\mathbf{Z}^v$, the l-th layer has the information losing rate $\gamma_l^v \geq 0$. If $\mathbf{S}$ is an oracle variable that contains and only contains multiple views' discriminative semantic information, and $\mathbf{H}^v$ is the t^v-th layer's features, we have minimizing the regularized loss $\mathcal{W}^{m,n}\mathcal{L}_{InfoNCE}^{m,n}(\mathbf{Z}^m, \mathbf{Z}^n) + \lambda \sum_v \mathcal{R}^v(\mathbf{X}^v, \mathbf{H}^v)$ is expected to obtain $I(\mathbf{S}; \mathbf{Z}^m; \mathbf{Z}^n) \leq \min\{I(\mathbf{S}; \mathbf{X}^m) \cdot \prod_{l=t^m+1}^{L^m}(1-\gamma_l^m), I(\mathbf{S}; \mathbf{X}^n) \cdot \prod_{l=t^n+1}^{L^n}(1-\gamma_l^n)\}$.*

Proof. We denote the hidden layers' features in encoders as $\mathbf{H}_{(1)}^v, \mathbf{H}_{(2)}^v, \ldots, \mathbf{H}_{(l)}^v, \ldots, \mathbf{H}_{(L^v)}^v$. Based on data processing inequality, we have

$$I(\mathbf{S}; \mathbf{X}^v) \geq I(\mathbf{S}; \mathbf{H}_{(1)}^v) \geq I(\mathbf{S}; \mathbf{H}_{(2)}^v) \geq \ldots I(\mathbf{S}; \mathbf{H}_{(l)}^v) \geq \ldots I(\mathbf{S}; \mathbf{H}_{(L^v)}^v) \geq I(\mathbf{S}; \mathbf{Z}^v). \tag{6}$$

Considering information losing, we have

$$I(\mathbf{S}; \mathbf{Z}^v) \leq I(\mathbf{S}; \mathbf{X}^v) \cdot \prod_{l=1}^{L^v}(1-\gamma_l^v). \tag{7}$$

According to Proposition 1 and Proposition 2, minimizing the regularized loss approximately becomes maximizing the following objective:

$$\mathcal{W}^{m,n} I(\mathbf{Z}^m; \mathbf{Z}^n) + \lambda \sum_{v=m,n} I(\mathbf{X}^v; \mathbf{H}^v), \tag{8}$$

where $\mathcal{W}^{m,n} > 0$ as two views ($v \in \{m, n\}$) are with positive class mutual information. The reconstruction regularization $I(\mathbf{X}^v; \mathbf{H}^v)$ makes $I(\mathbf{S}; \mathbf{X}^v) = I(\mathbf{S}; \mathbf{H}^v)$. Therefore, if $\mathbf{H}^v$ is the t^v-th layer's features (*i.e.*, $\mathbf{H}_{(t^v)}^v$ act as the regularized hidden features), we have

$$I(\mathbf{S}; \mathbf{Z}^v) \leq I(\mathbf{S}; \mathbf{X}^v) \cdot \prod_{l=t^v+1}^{L^v}(1-\gamma_l^v). \tag{9}$$

The contrastive loss leads to $\max I(\mathbf{Z}^m; \mathbf{Z}^n)$ which essentially explores the discriminative semantic information in $\mathbf{S}$. Given $I(\mathbf{S}; \mathbf{Z}^m; \mathbf{Z}^n) \leq \min\{I(\mathbf{S}; \mathbf{Z}^m), I(\mathbf{S}; \mathbf{Z}^n)\}$, as a result, we can obtain the mutual information across $\mathbf{S}$, $\mathbf{Z}^m$, and $\mathbf{Z}^n$ as follows:

$$I(\mathbf{S}; \mathbf{Z}^m; \mathbf{Z}^n) \leq \min\{I(\mathbf{S}; \mathbf{X}^m) \cdot \prod_{l=t^m+1}^{L^m}(1-\gamma_l^m), I(\mathbf{S}; \mathbf{X}^n) \cdot \prod_{l=t^n+1}^{L^n}(1-\gamma_l^n)\}. \tag{10}$$

□

In SEM, we have $\mathbf{Z}^v = g^v(\mathbf{H}^v; \Phi^v)$ where $\mathbf{H}^v$ and $\mathbf{Z}^v$ are two different variables. This design aims at separately maintaining different views' discriminative information by $\{\mathbf{H}^v\}_{v=1}^V$ and exploring their common semantic information by $\{\mathbf{Z}^v\}_{v=1}^V$. Contrastive learning on $\{\mathbf{Z}^v\}_{v=1}^V$ will capture the common semantics across multiple views induced by the contrastive loss, and discard other useless information in $\{\mathbf{H}^v\}_{v=1}^V$. In extreme cases, if we consider g^v as a smooth invertible transformation and we have the following theorem:

Theorem 3. *For any two views ($v \in \{m, n\}$) with positive class mutual information $I(\mathbf{y}^m; \mathbf{y}^n) = \delta$, $\delta > 0$, if g^v learned by contrastive learning is a smooth invertible transformation, minimizing the regularized loss $\mathcal{W}^{m,n}\mathcal{L}_{InfoNCE}^{m,n}(\mathbf{Z}^m, \mathbf{Z}^n) + \lambda \sum_{v=m,n} \mathcal{R}^v(\mathbf{X}^v, \mathbf{H}^v)$ will lead to a trade-off between $\max I(\mathbf{X}^m; \mathbf{X}^n; \mathbf{H}^m; \mathbf{H}^n)$ and $\max I(\mathbf{X}^m; \mathbf{H}^m) + I(\mathbf{X}^n; \mathbf{H}^n)$.*

Proof. According to Proposition 1 and Proposition 2, minimizing the regularized loss approximately becomes maximizing the following objective:

$$(e^{\delta/\log N} - 1) I(\mathbf{Z}^m; \mathbf{Z}^n) + \lambda I(\mathbf{X}^m; \mathbf{H}^m) + \lambda I(\mathbf{X}^n; \mathbf{H}^n). \tag{11}$$

If transformations g^m and g^n are smooth and invertible, the Jacobian determinant is $J_{\mathbf{Z}^m} = |\frac{\partial \mathbf{Z}^m}{\partial \mathbf{H}^m}|$ and $J_{\mathbf{Z}^n} = |\frac{\partial \mathbf{Z}^n}{\partial \mathbf{H}^n}|$, respectively. For the m-th and n-th views, we have

$$\begin{aligned} &p(\mathbf{h}^m, \mathbf{h}^n) = p(\mathbf{z}^m, \mathbf{z}^n) J_{\mathbf{Z}^m}(\mathbf{h}^m) J_{\mathbf{Z}^n}(\mathbf{h}^n), \\ &p(\mathbf{h}^m) = p(\mathbf{z}^m) J_{\mathbf{Z}^m}(\mathbf{h}^m), d\mathbf{z}^m = J_{\mathbf{Z}^m}(\mathbf{z}^m) d\mathbf{h}^m, \\ &p(\mathbf{h}^n) = p(\mathbf{z}^n) J_{\mathbf{Z}^n}(\mathbf{h}^n), d\mathbf{z}^n = J_{\mathbf{Z}^n}(\mathbf{z}^n) d\mathbf{h}^n. \end{aligned} \tag{12}$$

Then, we can obtain the invariance property of mutual information between $I(\mathbf{Z}^m;\mathbf{Z}^n)$ and $I(\mathbf{H}^m;\mathbf{H}^n)$ as follows:

$$\begin{aligned} I(\mathbf{Z}^m;\mathbf{Z}^n) &= \int\int p(\mathbf{z}^m,\mathbf{z}^n)\log\left(\frac{p(\mathbf{z}^m,\mathbf{z}^n)}{p(\mathbf{z}^m)p(\mathbf{z}^n)}\right)d\mathbf{z}^m d\mathbf{z}^n \\ &= \int\int \frac{p(\mathbf{h}^m,\mathbf{h}^n)}{J_{\mathbf{Z}^m}(\mathbf{h}^m)J_{\mathbf{Z}^n}(\mathbf{h}^n)}\log\left(\frac{\frac{p(\mathbf{h}^m,\mathbf{h}^n)}{J_{\mathbf{Z}^m}(\mathbf{h}^m)J_{\mathbf{Z}^n}(\mathbf{h}^n)}}{\frac{p(\mathbf{h}^m)p(\mathbf{h}^n)}{J_{\mathbf{Z}^m}(\mathbf{h}^m)J_{\mathbf{Z}^n}(\mathbf{h}^n)}}\right)J_{\mathbf{Z}^m}(\mathbf{z}^m)d\mathbf{h}^m J_{\mathbf{Z}^n}(\mathbf{z}^n)d\mathbf{h}^n \\ &= \int\int p(\mathbf{h}^m,\mathbf{h}^n)\log\left(\frac{p(\mathbf{h}^m,\mathbf{h}^n)}{p(\mathbf{h}^m)p(\mathbf{h}^n)}\right)d\mathbf{h}^m d\mathbf{h}^n \\ &= I(\mathbf{H}^m;\mathbf{H}^n). \end{aligned} \tag{13}$$

As a result, the optimization objective in Eq. (11) becomes

$$(e^{\delta/\log N}-1)I(\mathbf{H}^m;\mathbf{H}^n)+\lambda I(\mathbf{X}^m;\mathbf{H}^m)+\lambda I(\mathbf{X}^n;\mathbf{H}^n). \tag{14}$$

The mutual information $I(\mathbf{X}^m;\mathbf{X}^n)$ in data $\mathbf{X}^m$ and $\mathbf{X}^n$ is fixed, and the mutual information $I(\mathbf{H}^m;\mathbf{H}^n)$ changes due to variables $\mathbf{H}^m$ and $\mathbf{H}^n$. Maximizing $I(\mathbf{H}^m;\mathbf{H}^n)$ makes variables to access $I(\mathbf{X}^m;\mathbf{X}^n)$, while maximizing $I(\mathbf{X}^m;\mathbf{H}^m)+I(\mathbf{X}^n;\mathbf{H}^n)$ tends to maintain all information of view-specific data in variables. Since $I(\mathbf{X}^m;\mathbf{H}^m)\neq I(\mathbf{X}^m;\mathbf{X}^n;\mathbf{H}^m;\mathbf{H}^n)$, there is a trade-off controlled by λ, *i.e.*, maximizing to access the mutual information $I(\mathbf{X}^m;\mathbf{X}^n)$ between two view's data, or maximizing $I(\mathbf{X}^m;\mathbf{H}^m)+I(\mathbf{X}^n;\mathbf{H}^n)$ between variables and view-specific data. □

Typically, g^v will not be a smooth invertible transformation such that $\mathbf{H}^v$ and $\mathbf{Z}^v$ learn different information of data $\mathbf{X}^v$. As we all know, data $\mathbf{X}^v$ in different views usually contain useful discriminative information for common semantics as well as semantic-irrelevant information. We introduce the reconstruction regularization on $\mathbf{H}^v$ to avoid that $\mathbf{H}^v$ loses the useful discriminative information of data (here, $\mathbf{H}^v$ also maintains some semantic-irrelevant information due to the information reconstruction). Then, contrastive learning on $\mathbf{Z}^v$ can make $\mathbf{Z}^v$ access sufficient discriminative information from $\mathbf{H}^v$ to further explore the common semantics of multiple views. However, if the reconstruction regularization is punished on $\mathbf{Z}^v$, $\mathbf{Z}^v$ will also retain the semantic-irrelevant information which might disturb $\mathbf{Z}^v$ to explore the common semantics of multiple views. Therefore, the reconstruction objective of our SEM framework is built on $\mathbf{H}^v$ instead of $\mathbf{Z}^v$, for reducing the interference of semantic-irrelevant information to the contrastive learning performed on $\mathbf{Z}^v$.

Proposition 1. *Minimizing the weighted InfoNCE losses among multiple views' representations* $\sum_{m,n}\mathcal{W}^{m,n}\mathcal{L}^{m,n}_{InfoNCE}(\mathbf{Z}^m,\mathbf{Z}^n)$ *is equivalent to maximizing their weighted mutual information* $\sum_{m,n}\mathcal{W}^{m,n}I(\mathbf{Z}^m;\mathbf{Z}^n)$.

Proof. In this part, we leverage $d(\mathbf{z}_i^m,\mathbf{z}_i^n)$ to denote the cosine distance between $\mathbf{z}_i^m\in\mathbf{Z}^m$ and $\mathbf{z}_i^n\in\mathbf{Z}^n$. Then, based on the inequality in Lemma 1, we have:

$$-\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{d(\mathbf{z}_i^m,\mathbf{z}_i^n)/\tau}}{\sum_{j=1}^{N}e^{d(\mathbf{z}_i^m,\mathbf{z}_j^n)/\tau}}\geq\log N-I(\mathbf{Z}^m;\mathbf{Z}^n), \tag{15}$$

We rewrite the positive and negative pairs in InfoNCE loss and can obtain the following inequality:

$$\begin{aligned} &-\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{d(\mathbf{z}_i^m,\mathbf{z}_i^n)/\tau}}{\sum_{j=1}^{N}\sum_{v=m,n}e^{d(\mathbf{z}_i^m,\mathbf{z}_j^v)/\tau}} \\ \geq&-\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{d(\mathbf{z}_i^m,\mathbf{z}_i^n)/\tau}}{\sum_{j=1}^{N}e^{d(\mathbf{z}_i^m,\mathbf{z}_j^n)/\tau}} \\ \geq&\log N-I(\mathbf{Z}^m;\mathbf{Z}^n). \end{aligned} \tag{16}$$

Given the equations $I(\mathbf{Z}^m;\mathbf{Z}^n)=I(\mathbf{Z}^n;\mathbf{Z}^m)$ and $\mathcal{W}^{m,n}=\mathcal{W}^{n,m}$, we further have

$$\begin{aligned} \sum\nolimits_{m,n}\mathcal{W}^{m,n}\mathcal{L}^{m,n}_{InfoNCE}(\mathbf{Z}^m,\mathbf{Z}^n)&\geq\sum_{m=1}^{V}\sum_{n=1}^{V}(\log N-\mathcal{W}^{m,n}I(\mathbf{Z}^m;\mathbf{Z}^n)) \\ &=V^2\log N-2\sum_{m=1}^{V}\sum_{n=m}^{V}\mathcal{W}^{m,n}I(\mathbf{Z}^m;\mathbf{Z}^n). \end{aligned} \tag{17}$$

Therefore, $\min \sum_{m,n} \mathcal{W}^{m,n} \mathcal{L}^{m,n}_{InfoNCE}(\mathbf{Z}^m, \mathbf{Z}^n)$ is equivalent to $\max \sum_{m,n} \mathcal{W}^{m,n} I(\mathbf{Z}^m; \mathbf{Z}^n)$, *i.e.*, minimizing the weighted InfoNCE losses among multiple views' representations is equivalent to maximizing their weighted mutual information. □

The success of contrastive learning is often (not absolutely) attributable to the estimation of mutual information. The following Eq. (18) gives the relation between InfoNCE and mutual information, which also has been discussed by other forms in [1, 2, 3, 4, 5]. In this paper, We rewrite a proof to this inequality for the completeness of lemmas.

Lemma 1. *Let m and n denote two views, assuming $p(\mathbf{z}_i^m, \mathbf{z}_j^n) = p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)$ when $j \neq i$, we have the following inequality that give the relation between InfoNCE and mutual information:*

$$-\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(d(\mathbf{z}_i^m, \mathbf{z}_i^n)/\tau)}{\sum_{j=1}^{N} \exp(d(\mathbf{z}_i^m, \mathbf{z}_j^n)/\tau)} \geq \log N - I(\mathbf{Z}^m; \mathbf{Z}^n). \tag{18}$$

Proof. If $j \neq i$, $p(\mathbf{z}_j^n | \mathbf{z}_i^m) = \frac{p(\mathbf{z}_j^n, \mathbf{z}_i^m)}{p(\mathbf{z}_i^m)} = p(\mathbf{z}_j^n)$. Let $\mathcal{S}_i = \sum_{j=1}^{N} \frac{p(\mathbf{z}_i^m, \mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)}$, therefore, we have

$$\begin{aligned}
I(\mathbf{Z}^m; \mathbf{Z}^n) &= \sum_{i=1}^{N}\sum_{j=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_j^n) \log \frac{p(\mathbf{z}_i^m, \mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_j^n) \log \left(\frac{p(\mathbf{z}_i^m, \mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n) \cdot \mathcal{S}_i} \cdot \mathcal{S}_i \right) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_j^n) \log \frac{\frac{p(\mathbf{z}_i^m, \mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)}}{\mathcal{S}_i} + \sum_{i=1}^{N}\sum_{j=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_j^n) \log \mathcal{S}_i \\
&= \sum_{i=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_i^n) \log \frac{\frac{p(\mathbf{z}_i^m, \mathbf{z}_i^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_i^n)}}{\mathcal{S}_i} + \sum_{i=1}^{N}\sum_{j\neq i} p(\mathbf{z}_i^m, \mathbf{z}_j^n) \log \frac{\frac{p(\mathbf{z}_i^m, \mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)}}{\mathcal{S}_i} \\
&+ \sum_{i=1}^{N}\sum_{j=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_j^n) \log \mathcal{S}_i. \\
&= \sum_{i=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_i^n) \log \frac{\frac{p(\mathbf{z}_i^m, \mathbf{z}_i^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_i^n)}}{\mathcal{S}_i} + \sum_{i=1}^{N}\sum_{j\neq i} p(\mathbf{z}_i^m, \mathbf{z}_j^n) \log \frac{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)} \\
&+ \sum_{i=1}^{N}\sum_{j=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_j^n) \log \mathcal{S}_i - \sum_{i=1}^{N}\sum_{j\neq i} p(\mathbf{z}_i^m, \mathbf{z}_j^n) \log \mathcal{S}_i \\
&= \sum_{i=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_i^n) \log \frac{\frac{p(\mathbf{z}_i^m, \mathbf{z}_i^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_i^n)}}{\mathcal{S}_i} + \sum_{i=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_i^n) \log \mathcal{S}_i.
\end{aligned} \tag{19}$$

Since positive pairs are correlated, we have the estimate: $p(\mathbf{z}_i^m, \mathbf{z}_i^n) \geq p(\mathbf{z}_i^m)p(\mathbf{z}_i^n)$. Therefore, the following inequality holds:

$$\begin{aligned}
\log \mathcal{S}_i &= \log \left(\sum_{j=1}^{N} \frac{p(\mathbf{z}_i^m, \mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)} \right) \\
&= \log \left(\frac{p(\mathbf{z}_i^m, \mathbf{z}_i^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_i^n)} + \sum_{j\neq i} \frac{p(\mathbf{z}_i^m, \mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)} \right) \\
&= \log \left(N + \frac{p(\mathbf{z}_i^m, \mathbf{z}_i^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_i^n)} - 1 \right) \\
&\geq \log N.
\end{aligned} \tag{20}$$

According to Lemma 2 and Eq. (20), we assume that there exists a constant $\delta \in (0,1)$ such that $p(\mathbf{z}_i^m|\mathbf{z}_i^n) \geq \delta, i = 1, 2, \cdots, N$ holds. With the estimation [1, 3], *i.e.*, $p(\mathbf{z}_i^n) \approx \frac{1}{N}, i = 1, 2, \cdots, N$, the following inequality holds:

$$\begin{aligned} I(\mathbf{Z}^m; \mathbf{Z}^n) &= \sum_{i=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_i^n) \log \frac{\frac{p(\mathbf{z}_i^m, \mathbf{z}_i^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_i^n)}}{\mathcal{S}_i} + \sum_{i=1}^{N} p(\mathbf{z}_i^m, \mathbf{z}_i^n) \log \mathcal{S}_i \\ &\approx \sum_{i=1}^{N} \frac{1}{N} p(\mathbf{z}_i^m|\mathbf{z}_i^n) \log \frac{\frac{p(\mathbf{z}_i^m, \mathbf{z}_i^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_i^n)}}{\mathcal{S}_i} + \sum_{i=1}^{N} \frac{1}{N} p(\mathbf{z}_i^m|\mathbf{z}_i^n) \log \mathcal{S}_i \\ &\geq \delta \left(\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(d(\mathbf{z}_i^m, \mathbf{z}_i^n)/\tau)}{\sum_{j=1}^{N} \exp(d(\mathbf{z}_i^m, \mathbf{z}_j^n)/\tau)} + \log N \right). \end{aligned} \tag{21}$$

Furthermore, we have

$$-\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(d(\mathbf{z}_i^m, \mathbf{z}_i^n)/\tau)}{\sum_{j=1}^{N} \exp(d(\mathbf{z}_i^m, \mathbf{z}_j^n)/\tau)} \geq \log N - \frac{1}{\delta} I(\mathbf{Z}^m; \mathbf{Z}^n). \tag{22}$$

Consequently, when the constant $\delta \approx 1$ (*i.e.*, the positive pairs are approximate to be correlated), Eq. (18) holds. □

According to [1], Eq. (22) is more precise when N is larger. Minimizing the left part of Eq. (22) is equivalent to maximizing the mutual information $I(\mathbf{Z}^m; \mathbf{Z}^n)$. Note that this bound is weak as there exists approximation about mutual information [6].

Lemma 2. *The optimal value of* $\exp(d(\mathbf{z}_i^m, \mathbf{z}_j^n)/\tau)$ *is proportional to the ratio of* $p(\mathbf{z}_i^m, \mathbf{z}_j^n)$ *to* $p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)$*, i.e.,* $\exp(d(\mathbf{z}_i^m, \mathbf{z}_j^n)/\tau) \propto \frac{p(\mathbf{z}_i^m, \mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)}$.

Proof. We consider the following formulation:

$$-\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(d(\mathbf{z}_i^m, \mathbf{z}_i^n)/\tau)}{\sum_{j=1}^{N} \exp(d(\mathbf{z}_i^m, \mathbf{z}_j^n)/\tau)}. \tag{23}$$

Eq. (23) can be regarded as a cross-entropy loss. As a result, minimizing this loss is equivalent to solving a binary classification problem, namely, classifying the given pairs into positive or negative pairs. We let $\{\mathbf{z}_i^m, \mathbf{z}_i^n\}$ denote the positive pairs and $\{\mathbf{z}_i^m, \mathbf{z}_j^n\}_{j \neq i}$ denote the negative pairs. For each given pairs $\{\mathbf{z}_i^m, \mathbf{z}_j^n\}_{i,j=1}^N$, we let $p(\mathbf{z}_i^n|\{\mathbf{z}_1^n, \cdots, \mathbf{z}_N^n\}, \mathbf{z}_i^m)$ denote the predicted probability of finding $\mathbf{z}_i^n$ from $\{\mathbf{z}_1^n, \cdots, \mathbf{z}_N^n\}$ to form positive pairs $\{\mathbf{z}_i^m, \mathbf{z}_i^n\}$. $p(\mathbf{z}_i^m, \mathbf{z}_j^n)$, $p(\mathbf{z}_i^m)$, and $p(\mathbf{z}_j^n)$ denote the joint probability and marginal probabilities of $\mathbf{z}_i^m$ and $\mathbf{z}_j^n$. Then, the optimal value of $p(\mathbf{z}_i^n|\{\mathbf{z}_1^n, \cdots, \mathbf{z}_N^n\}, \mathbf{z}_i^m)$ is:

$$\begin{aligned} p(\mathbf{z}_i^n|\{\mathbf{z}_1^n, \cdots, \mathbf{z}_N^n\}, \mathbf{z}_i^m) &= \frac{p(\mathbf{z}_i^n|\mathbf{z}_i^m) \prod_{l \neq i} p(\mathbf{z}_l^n)}{\sum_{j=1}^{N} p(\mathbf{z}_j^n|\mathbf{z}_i^m) \prod_{l \neq j} p(\mathbf{z}_l^n)} \\ &= \frac{\frac{p(\mathbf{z}_i^n|\mathbf{z}_i^m)}{p(\mathbf{z}_i^n)}}{\sum_{j=1}^{N} \frac{p(\mathbf{z}_j^n|\mathbf{z}_i^m)}{p(\mathbf{z}_j^n)}} \\ &= \frac{\frac{p(\mathbf{z}_i^m, \mathbf{z}_i^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_i^n)}}{\sum_{j=1}^{N} \frac{p(\mathbf{z}_i^m, \mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)}}. \end{aligned} \tag{24}$$

The corresponding cross-entropy loss is:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log p(\mathbf{z}_i^n|\{\mathbf{z}_1^n, \cdots, \mathbf{z}_N^n\}, \mathbf{z}_i^m) = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\frac{p(\mathbf{z}_i^m, \mathbf{z}_i^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_i^n)}}{\sum_{j=1}^{N} \frac{p(\mathbf{z}_i^m, \mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)}}. \tag{25}$$

Comparing Eq. (25) with Eq. (23), we can find $\exp(d(\mathbf{z}_i^m, \mathbf{z}_j^n)/\tau) \propto \frac{p(\mathbf{z}_i^m, \mathbf{z}_j^n)}{p(\mathbf{z}_i^m)p(\mathbf{z}_j^n)}$. □

Proposition 2. ($\max I(\mathbf{X}^v;\mathbf{H}^v)$ [7]) *Combining with Monte Carlo sampling, minimizing the reconstruction loss between raw data and reconstructed data* $\left\|\mathbf{X}^v - f_-^v(\mathbf{H}^v)\right\|_F^2$ *is approximate to maximizing the mutual information between raw data and their hidden features* $I(\mathbf{X}^v;\mathbf{H}^v)$.

Proof. For the v-th view, we let $\mathbf{x}^v$ and $\mathbf{h}^v$ denote the points in the space of raw data $\mathbf{X}^v$ and in the space of hidden features $\mathbf{H}^v$, respectively. According to the definition, the mutual information between $\mathbf{X}^v$ and $\mathbf{H}^v$ can be formulated as

$$I(\mathbf{X}^v;\mathbf{H}^v) = \int_{\mathbf{h}^v}\int_{\mathbf{x}^v} p(\mathbf{x}^v,\mathbf{h}^v)\log\left(\frac{p(\mathbf{x}^v|\mathbf{h}^v)}{p(\mathbf{x}^v)}\right)d\mathbf{x}^v d\mathbf{h}^v. \tag{26}$$

The decoder network achieves the approximation $q(\mathbf{x}^v|\mathbf{h}^v)$ of the true posterior $p(\mathbf{x}^v|\mathbf{h}^v)$. Based on the non-negative property of Kullback-Leibler divergence (D_{KL}), we have

$$\begin{aligned}
&\int_{\mathbf{x}^v} p(\mathbf{x}^v|\mathbf{h}^v)\log\left(\frac{p(\mathbf{x}^v|\mathbf{h}^v)}{q(\mathbf{x}^v|\mathbf{h}^v)}\right)d\mathbf{x}^v = D_{KL}[p(\mathbf{x}^v|\mathbf{h}^v)||q(\mathbf{x}^v|\mathbf{h}^v)] \geq 0\\
\Rightarrow &\int_{\mathbf{x}^v} p(\mathbf{x}^v|\mathbf{h}^v)\log\left(p(\mathbf{x}^v|\mathbf{h}^v)\right)d\mathbf{x}^v \geq \int_{\mathbf{x}^v} p(\mathbf{x}^v|\mathbf{h}^v)\log\left(q(\mathbf{x}^v|\mathbf{h}^v)\right)d\mathbf{x}^v\\
\Rightarrow &\int_{\mathbf{h}^v} p(\mathbf{h}^v)d\mathbf{h}^v\int_{\mathbf{x}^v} p(\mathbf{x}^v|\mathbf{h}^v)\log\left(p(\mathbf{x}^v|\mathbf{h}^v)\right)d\mathbf{x}^v\\
&\geq \int_{\mathbf{h}^v} p(\mathbf{h}^v)d\mathbf{h}^v\int_{\mathbf{x}^v} p(\mathbf{x}^v|\mathbf{h}^v)\log\left(q(\mathbf{x}^v|\mathbf{h}^v)\right)d\mathbf{x}^v\\
\Rightarrow &\int_{\mathbf{h}^v}\int_{\mathbf{x}^v} p(\mathbf{x}^v,\mathbf{h}^v)\log\left(p(\mathbf{x}^v|\mathbf{h}^v)\right)d\mathbf{x}^v d\mathbf{h}^v\\
&\geq \int_{\mathbf{h}^v}\int_{\mathbf{x}^v} p(\mathbf{x}^v,\mathbf{h}^v)\log\left(q(\mathbf{x}^v|\mathbf{h}^v)\right)d\mathbf{x}^v d\mathbf{h}^v\\
\Rightarrow &\int_{\mathbf{h}^v}\int_{\mathbf{x}^v} p(\mathbf{x}^v,\mathbf{h}^v)\log\left(\frac{p(\mathbf{x}^v|\mathbf{h}^v)}{p(\mathbf{x}^v)}\right)d\mathbf{x}^v d\mathbf{h}^v\\
&\geq \int_{\mathbf{h}^v}\int_{\mathbf{x}^v} p(\mathbf{x}^v,\mathbf{h}^v)\log\left(\frac{q(\mathbf{x}^v|\mathbf{h}^v)}{p(\mathbf{x}^v)}\right)d\mathbf{x}^v d\mathbf{h}^v\\
\Rightarrow &I(\mathbf{X}^v;\mathbf{H}^v) \geq \int_{\mathbf{h}^v}\int_{\mathbf{x}^v} p(\mathbf{x}^v,\mathbf{h}^v)\log\left(\frac{q(\mathbf{x}^v|\mathbf{h}^v)}{p(\mathbf{x}^v)}\right)d\mathbf{x}^v d\mathbf{h}^v.
\end{aligned} \tag{27}$$

Considering $-\int_{\mathbf{h}^v}\int_{\mathbf{x}^v} p(\mathbf{x}^v,\mathbf{h}^v)\log\left(p(\mathbf{x}^v)\right)d\mathbf{x}^v d\mathbf{h}^v \geq 0$, we further have

$$\begin{aligned}
I(\mathbf{X}^v;\mathbf{H}^v) &\geq \int_{\mathbf{h}^v}\int_{\mathbf{x}^v} p(\mathbf{x}^v,\mathbf{h}^v)\log\left(q(\mathbf{x}^v|\mathbf{h}^v)\right)d\mathbf{x}^v d\mathbf{h}^v\\
&\quad - \int_{\mathbf{h}^v}\int_{\mathbf{x}^v} p(\mathbf{x}^v,\mathbf{h}^v)\log\left(p(\mathbf{x}^v)\right)d\mathbf{x}^v d\mathbf{h}^v\\
\Rightarrow I(\mathbf{X}^v;\mathbf{H}^v) &\geq \int_{\mathbf{h}^v}\int_{\mathbf{x}^v} p(\mathbf{x}^v,\mathbf{h}^v)\log\left(q(\mathbf{x}^v|\mathbf{h}^v)\right)d\mathbf{x}^v d\mathbf{h}^v\\
\Rightarrow I(\mathbf{X}^v;\mathbf{H}^v) &\geq \int_{\mathbf{x}^v} p(\mathbf{x}^v)d\mathbf{x}^v\int_{\mathbf{h}^v} p(\mathbf{h}^v|\mathbf{x}^v)\log\left(q(\mathbf{x}^v|\mathbf{h}^v)\right)d\mathbf{h}^v.
\end{aligned} \tag{28}$$

Based on Monte Carlo sampling method [8, 7] on $\mathbf{x}_i^v \in \mathbf{X}^v$, we obtain

$$\begin{aligned}
&\int_{\mathbf{x}^v} p(\mathbf{x}^v)d\mathbf{x}^v\int_{\mathbf{h}^v} p(\mathbf{h}^v|\mathbf{x}^v)\log\left(q(\mathbf{x}^v|\mathbf{h}^v)\right)d\mathbf{h}^v\\
=&\frac{1}{N}\sum_{i=1}^{N}\int_{\mathbf{h}^v} p(\mathbf{h}^v|\mathbf{x}_i^v)\log\left(q(\mathbf{x}_i^v|\mathbf{h}^v)\right)d\mathbf{h}^v\\
=&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{p(\mathbf{h}^v|\mathbf{x}_i^v)}[\log\left(q(\mathbf{x}_i^v|\mathbf{h}^v)\right)],
\end{aligned} \tag{29}$$

where $p(\mathbf{h}^v|\mathbf{x}_i^v)$ and $q(\mathbf{x}_i^v|\mathbf{h}^v)$ could be treated as the encoder f^v and decoder f_-^v processes of $\mathbf{x}_i^v$, respectively. There is no harm in supposing that $q(.)$ follows Gaussion distribution [7]. Then, the approximate posterior $q(\mathbf{x}_i^v|\mathbf{h}^v)$ can be formulated as

$$q(\mathbf{x}_i^v|\mathbf{h}^v) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left\|\mathbf{x}_i^v - f_-^v(\mathbf{h}^v)\right\|_2^2}{\sigma^2}\right). \tag{30}$$

As a result, we have the following inequality:

$$I(\mathbf{X}^v;\mathbf{H}^v) \geq \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{p(\mathbf{h}^v|\mathbf{x}_i^v)}\left[-\log\left(\sqrt{2\pi}\sigma\right) - \frac{\left\|\mathbf{x}_i^v - f_-^v(\mathbf{h}^v)\right\|_2^2}{\sigma^2}\right]. \tag{31}$$

Therefore, minimizing $\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{p(\mathbf{h}^v|\mathbf{x}_i^v)}\left[\left\|\mathbf{x}_i^v - f_-^v(\mathbf{h}^v)\right\|_2^2\right]$ is approximate to maximizing $I(\mathbf{X}^v;\mathbf{H}^v)$. If we continue to simplify $\mathbf{h}_i^v \in \mathbf{H}^v$ with Monte Carlo sampling method, we further have

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{p(\mathbf{h}^v|\mathbf{x}_i^v)}\left[\left\|\mathbf{x}_i^v - f_-^v(\mathbf{h}^v)\right\|_2^2\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}\left[\left\|\mathbf{x}_i^v - f_-^v(\mathbf{h}_{i(j)}^v)\right\|_2^2\right]. \tag{32}$$

Since $\mathbf{h}_i^v$ can be the only one output of $\mathbf{x}_i^v$ by decoder network [7] (*i.e.*, $M = 1$), we could obtain that minimizing the reconstruction loss $\left\|\mathbf{X}^v - f_-^v(\mathbf{H}^v)\right\|_F^2$ is approximate to maximizing $I(\mathbf{X}^v;\mathbf{H}^v)$. □

Complexity analysis Letting N, n, E represent the data size, batch size, and total training epochs, respectively, the computation of loss functions and the update of model parameters are with mini-batch manner. Their time complexity is determined by the batch size n and the total training epochs E. Since $n \ll N$ holds, the complexity would be $O(E)$. Letting V represent the number of views, h_v and z denote the dimensionality of $\mathbf{H}^v$ and $\mathbf{Z}^v$ of the v-th view, respectively. Step size S denotes the number of training epochs after each update of weights. In terms of weighting strategy $\mathcal{W}_{MMD}$, for reducing the complexity of MMD, we can leverage partial instead of whole samples to update $\{\mathcal{W}^{m,n}\}_{m,n=1}^{V}$. For example, we can randomly pick up $\hat{n}$ samples ($\hat{n} \ll N$) to compute MMD and the complexity is just $O(\hat{n}^2)$. For weighting strategy $\mathcal{W}_{CMI}$, the computation of $\{\mathcal{W}^{m,n}\}_{m,n=1}^{V}$ needs $V \times N$ representations from all views to obtain K-Means clustering results and its total time complexity is $O(h_v VN) + O(zVNE/S)$, which is linear to N. When N is too large, we can apply mini-batch K-Means to reduce the complexity of CMI weighting strategy.

Appendix B Experimental Settings

Table 1: Description for abbreviation

Abbr.	Description
InfoNCE	Info noise contrastive estimation
SIFT	Scale-invariant feature transform
STIP	Space-time interest points
MFCC	Mel-frequency cepstral coefficents
CENTRIST	Census transform histogram
HOG	Histogram of oriented gradient
LBP	Local binary pattern

The models of all methods are implemented with PyTorch [9] platform and tested on the same device with a NVIDIA GeForce RTX 3090 GPU (24.0GB caches) and a 11th Gen Intel(R) Core(TM) i5-11600KF @ 3.90GHz CPU (64.0GB RAM). For fair comparison, all methods adopt the similar architecture of neural networks following previous work [10, 11]. For our SEM, the encoder network can be denoted as $\mathbf{X}^v \to 500 \to 500 \to 2000 \to \mathbf{H}^v \to \mathbf{Z}^v$ and the decoder is reversed $\mathbf{H}^v \to 2000 \to 500 \to 500 \to \hat{\mathbf{X}}^v$. In this architecture, the penultimate layer of encoder networks is recorded as the hidden features $\mathbf{H}^v$. For all views, the dimension of hidden features $\mathbf{H}^v$ and contrastive representations $\mathbf{Z}^v$ are set to 512 and 128, respectively. Activation function is ReLU [12] and optimizer adopts Adam [13]. For all used datasets, the learning rate is fixed to 0.0003 and the

hyper-parameter λ is fixed to 1. Table 2 shows the network training details on different datasets. In our experiments, as computing MMD has high complexity, we select first 2000 samples for avoiding out-of-memory when data size is large. The noise of denoising autoencoder is random Gaussian noise. The mask rate of masked autoencoder is set to 30%. For the CMI weighting strategy, the cluster number of K-Means algorithm is pre-defined to the truth class number of a dataset. For the MMD weighting strategy, the bandwith and number of kernels are set to 4 for all datasets used in this paper. Since our work does not focus on specific contrastive losses, we adopt the fixed parameter settings of InfoNCE/PSCL/RINCE as shown in Table 3 for all experiments. Moreover, the batch size that will affect the number of negative pairs is also fixed to 256.

Table 2: Network training details on different datasets

	Pre-training Epoch	CL Epoch	Input dimensions of different views	dimension of $\mathbf{H}^v$	dimension of $\mathbf{Z}^v$
DHA	100	300	110/6144	512	128
CCV	100	200	5000/5000/4000	512	128
NUSWIDE	100	100	64/225/144/73/128	512	128
Caltech	100	400	48/40/254/1984/512/928	512	128
YoutubeVideo	100	25	512/647/838	512	128

Table 3: Parameter setting in contrastive losses

	Parameters
InfoNCE	$\tau = 1.0$
PSCL	$r = 3.0$
RINCE	$\tau = 0.5, \alpha = 0.001, q = 0.5$

Appendix C Additional Experiments

Figure 1 and 2 show the linear classification performance on NUSWIDE and Caltech datasets. We do not report the results on YoutubeVideo as this large-scale dataset is beyond the usable range of SVM.

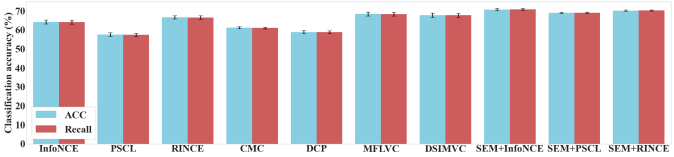


Figure 1: Classification performance on NUSWIDE.

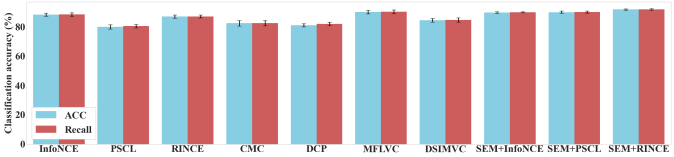


Figure 2: Classification performance on Caltech.

Table 4 reports the time consumption of SEM with three options of weight strategy on five datasets, where the contrastive loss and reconstruction term are fixed to $\mathcal{L}_{InfoNCE}$ and $\mathcal{R}_{AE}$. On CCV, NUSWIDE, and YoutubeVideo, as MMD has high complexity, we select first 2,000 samples to compute weights for avoiding out-of-memory. In this setting, we observe that SEM w/ $\mathcal{W}_{JSD}$ is the fastest variant among the three variants as the computation of JSD is the simplest. Generally, SEM w/ $\mathcal{W}_{CMI}$ is faster than SEM w/ $\mathcal{W}_{MMD}$ even if the MMD is computed on partial data.

Table 5 reports the results of ablation experiments on SEM with different options of weight strategy on five datasets. Table 6 reports the results of ablation experiments on SEM with different options of reconstruction term on five datasets, where $\mathcal{R}_{AE/DAE/MAE}$ w/o SEM denote the performance on representations learned by AE/DAE/MAE models without SEM framework. We can observe that SEM w/ $\mathcal{R}_{AE/DAE/MAE}$ achieve significant improvements over $\mathcal{R}_{AE/DAE/MAE}$ w/o SEM.

Table 4: Time consumption (seconds) of SEM with different options of weight strategy

Variants	DHA	CCV	NUSWIDE	Caltech	YoutubeVideo
SEM w/ $\mathcal{W}_{CMI}$	38	984	783	533	1990
SEM w/ $\mathcal{W}_{JSD}$	25	556	389	396	1938
SEM w/ $\mathcal{W}_{MMD}$	28	833	1144	775	2248

Table 5: Clustering performance on SEM with different options of $\mathcal{W}$

	DHA		CCV		NUSWIDE		Caltech		YoutubeVideo	
Variants	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
SEM w/o $\mathcal{W}$	71.29	79.77	33.50	33.01	61.90	33.82	77.71	68.68	20.96	20.82
SEM w/ $\mathcal{W}_{CMI}$	80.87	84.10	39.35	35.50	60.37	34.92	87.17	80.33	31.25	31.12
SEM w/ $\mathcal{W}_{JSD}$	80.53	83.75	35.59	33.45	62.96	34.61	85.50	77.16	21.76	21.49
SEM w/ $\mathcal{W}_{MMD}$	84.40	86.22	33.89	34.13	61.39	32.75	85.67	77.36	27.50	28.47

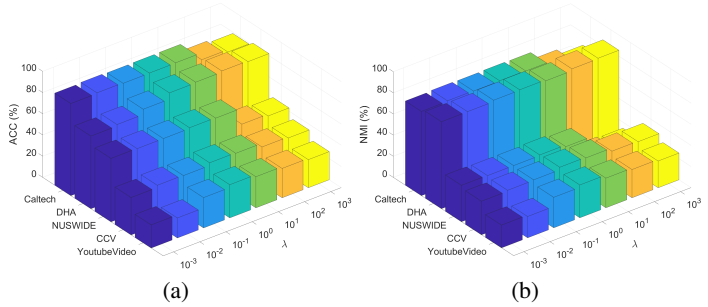
Table 6: Clustering performance on SEM with different options of $\mathcal{R}$

	DHA		CCV		NUSWIDE		Caltech		YoutubeVideo	
Variants	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
$\mathcal{R}_{AE}$ w/o SEM	69.15	78.43	14.29	11.39	38.70	13.60	86.00	76.43	20.03	19.55
$\mathcal{R}_{DAE}$ w/o SEM	70.39	78.87	12.67	9.58	39.54	15.08	86.43	77.47	21.73	21.49
$\mathcal{R}_{MAE}$ w/o SEM	69.98	77.10	14.62	11.66	35.84	14.54	86.21	77.08	22.78	21.95
SEM w/o $\mathcal{R}$	60.45	74.11	28.72	26.53	57.74	26.62	79.42	69.78	32.69	32.57
SEM w/ $\mathcal{R}_{AE}$	80.87	84.10	39.35	35.50	60.37	34.92	87.17	80.33	31.25	31.12
SEM w/ $\mathcal{R}_{DAE}$	81.50	83.49	38.42	33.62	59.54	33.62	86.57	79.12	38.78	36.70
SEM w/ $\mathcal{R}_{MAE}$	83.02	84.44	39.48	35.79	60.94	36.24	86.71	78.03	33.26	33.04

Table 7 reports the experiments on SEM with different sum manner of contrastive losses (where the combination of $\mathcal{L}_{InfoNCE}$+$\mathcal{W}_{CMI}$+$\mathcal{R}_{AE}$ is taken). We observe that $\sum_m^V \sum_{n=m+1}^V \mathcal{L}_{CL}^{m,n}$ performs worse than $\sum_m^V \sum_n^V \mathcal{L}_{CL}^{m,n}$. This might be because the latter (*i.e.*, $\sum_{m,n} \mathcal{L}_{CL}^{m,n}$ in the paper) pairs negative samples for both $\mathbf{z}_i^m$ and $\mathbf{z}_i^n$ (*e.g.*, $\{\mathbf{z}_i^m, \mathbf{z}_j^v\}_{j\neq i}^{v=m,n}$ and $\{\mathbf{z}_i^n, \mathbf{z}_j^v\}_{j\neq i}^{v=n,m}$), which can access more comprehensive negative sample pairs for contrastive learning than the former.

Table 7: Clustering performance on SEM with different sum manner of contrastive losses

	DHA		CCV		NUSWIDE		Caltech		YoutubeVideo	
Variants	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
$\sum_m^V \sum_{n=m+1}^V \mathcal{L}_{CL}^{m,n}$	72.04	78.38	27.95	29.92	58.28	29.74	86.54	78.35	21.47	21.82
$\sum_m^V \sum_n^V \mathcal{L}_{CL}^{m,n}$	80.87	84.10	39.35	35.50	60.37	34.92	87.17	80.33	31.25	31.12

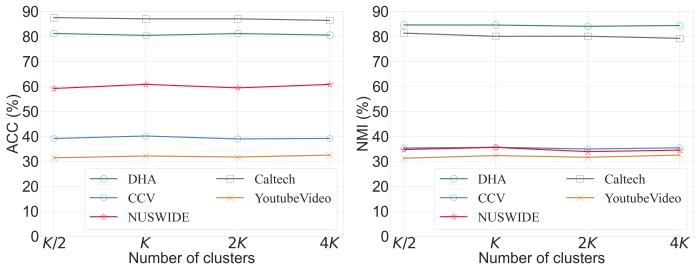


(a) (b)

Figure 3: (a) ACC $vs.$ λ. (b) NMI $vs.$ λ.

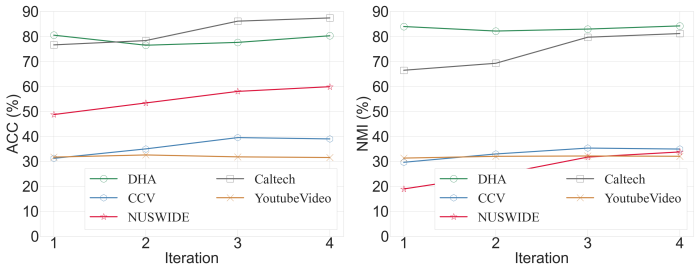
Parameter analysis Since different datasets have different levels of reconstruction errors, the trade-off coefficient λ is introduced to balance contrastive learning and information recovery in our SEM framework. In Figure 3, we change λ within the range of $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}]$

and report the clustering accuracy on the learned representations. The results indicate that SEM framework is not sensitive to λ in $[10^{-1}, 10^{1}]$. For all our experiments, λ is consistently set to 1. Additionally, we investigate the effect of cluster number when the weight strategy of SEM framework is selected as $\mathcal{W}_{CMI}$ that needs to pre-define the cluster number when applying K-Means. When computing the class mutual information, as shown in Figure 4, we change the number of clusters within the range of $[K/2, K, 2K, 4K]$ where K denotes the truth class number of datasets. The results demonstrate that SEM with $\mathcal{W}_{CMI}$ is not sensitive to the choices of cluster number.



(a) (b)

Figure 4: (a) ACC $vs.$ K. (b) ACC $vs.$ K.



(a) (b)

Figure 5: (a) ACC $vs.$ Iterative times of updating weights. (b) NMI $vs.$ Iterative times of updating weights.

In experiments, the times of updating weights during whole training is E/S, where E is total training epochs and S is the step size (the number of training epochs after each update of weights). In Figure 5, we fix S after each update of weights and record the clustering accuracy on the learned representations during the iterative times of updating weights in SEM framework (here, an iteration means the one time of updating weights). We observe that only one time of updating weights is enough for some datasets. Usually, the effect of multi-view contrastive learning is gradually improved with the increase of the times of updating weights for some datasets. In our experiments, we fix the times of updating weights on DHA/YoutubeVideo to 1, and fix those on CCV/NUSWIDE/Caltech to 4.

Appendix D Social Impacts and Limitations

In multi-view contrastive learning, it might be promising to take the representation degeneration into account, especially in unsupervised environments where the qualities of different views captured by various sensors cannot be guaranteed, *e.g.*, the views from some sensors in real-world application scenarios (such as in animal protection and automatic pilot) are faulty or not applicable, and thus bring semantic-irrelevant information. Additionally, our work proposed a machine learning algorithm to make contrastive learning more practical in the field of multi-view learning. This research is not expected to introduce new negative societal impacts beyond what is already known. Conceptually, the limitation of the self-weighting strategy is that it is more effective when there are over two views. When there are only two views, the self-weighted contrastive learning transforms into traditional contrastive learning but with reconstruction regularization. Therefore, one of our future work is to extend the view-level weighting of our proposed framework to sample-level weighting.

References

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

[2] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, pages 776–794, 2020.

[3] Huasong Zhong, Chong Chen, Zhongming Jin, and Xian-Sheng Hua. Deep robust clustering by contrastive learning. *arXiv preprint arXiv:2008.03030*, 2020.

[4] Hanwei Wu, Ather Gattami, and Markus Flierl. Conditional mutual information-based contrastive loss for financial time series forecasting. In *Proceedings of the ACM International Conference on AI in Finance*, pages 1–7, 2020.

[5] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939, 2020.

[6] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 875–884, 2020.

[7] Lei Zhang, Lele Fu, Tong Wang, Chuan Chen, and Chuanfu Zhang. Mutual information-driven multi-view clustering. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 3268–3277, 2023.

[8] Alexander Shapiro. Monte carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425, 2003.

[9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8024–8035, 2019.

[10] Huayi Tang and Yong Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International Conference on Machine Learning*, pages 21090–21110, 2022.

[11] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16051–16060, 2022.

[12] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.