Appendix for A Novel Approach for Effective Multi-View Clustering with Information-Theoretic Perspective

Anonymous Submission

A Formulas about Mutual Information

In this section we introduce several formulas about mutual information that are used in this work.

(P1) Positivity:

$$I(x; y) \geq 0, I(x; y|z) \geq 0$$

(P2) Chain rule:

$$I(xy; z) = I(y; z) + I(x; z|y)$$

(P3) Chain rule (Multivariate Mutual Information):

$$I(x; y; z) = I(y; z) - I(y; z|x)$$

B Theorems and Loss Computation

B.1 Decomposition of $I(x^i; z^i)$

Theorem B.1. *Let z^i be a representation of x^i, then the mutual information between x^i and z^i can be decomposed as:*

$$I(x^i, z^i) = I(z^i, x^i|x^j) + I(x^j, z^i).$$

Proof. Hypothesis (H1): z^i is a representation of x^i: $I(x^j; z^i|x^i) = 0$.
Based on the Hypothesis (H1) and Chain rule (P3), then we have

$$\begin{aligned} I(x^i; z^i) &\overset{(P3)}{=} I(x^i; z^i|x^j) + I(x^i; z^i; x^j) \\ &\overset{(P3)}{=} I(x^j; z^i|x^j) + I(x^j; z^i) - I(x^j; z^i|x^i) \\ &\overset{(H1)}{=} I(x^i; z^i|x^j) + I(x^j; z^i) \end{aligned}$$

□

B.2 Multi-View Redundancy and Sufficiency [2]

Theorem B.2. *Let x^i, x^j, and y be random variables with joint distribution $p(x^i, x^j, y)$. Let z^i be a representation of x^i, then:*

$$I(x^i; y|z^i) \leq I(x^i; x^j|z^i) + I(x^i; y|x^j).$$

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Proof. Hypothesis (H1): z^i is a representation of x^i: $I(y; z^i|x^j x^i) = 0$.
Based on the Hypothesis (H1), Positivity (P1) and Chain rule (P3), we have

$$\begin{aligned}
I(x^i; y|z^i) &\overset{(P3)}{=} I(x^i; y|z^i x^j) + I(x^i; x^j; y|z^i) \\
&\overset{(P3)}{=} I(x^i; y|x^j) - I(x^i; y; z^i|x^j) + I(x^i; x^j; y|z^i) \\
&\overset{(P3)}{=} I(x^i; y|x^j) - I(y; z^i|x^j) + I(y; z^i|x^j x^i) + I(x^i; x^j; y|z^i) \\
&\overset{(P1)}{\leq} I(x^i; y|x^j) + I(y; z^i|x^j x^i) + I(x^i; x^j; y|z^i) \overset{(H1)}{=} I(x^i; y|x^j) + I(x^i; x^j; y|z^i) \\
&\overset{(P3)}{=} I(x^i; y|x^j) + I(x^i; x^j|z^i) - I(x^i; x^j|z^i y) \\
&\overset{(P1)}{\leq} I(x^i; y|x^j) + I(x^i; x^j|z^i)
\end{aligned}$$

□

Theorem B.3. *Let x^i be a redundant view with respect to x^j for y. Any representation z^i of x^i that is sufficient for x^j is also sufficient for y, i.e.,*

$$I(x^i; x^j|z^i) = 0 \rightarrow I(x^i; y|z^i) = 0.$$

Proof. Hypotheses (H1): z^i is a representation of x^i: $I(y; z^i|x^i) = 0$.
(H2): x^i is redundant with respect to x^j for y: $I(y; x^i|x^j) = 0$.
Using the results from Theorem B.2 $I(x^i; y|z^i) \overset{(Th.B.2)}{\leq} I(x^i; y|x^j) + I(x^i; x^j|z^i) \overset{(H2)}{=} I(x^i; x^j|z^i))$, we have
$I(x^i; x^j|z^i) = 0 \rightarrow I(x^i; y|z^i) = 0$. □

Theorem B.4. *Let x^i, x^j and y be random variables with distribution $p(x^i, x^j, y)$. Let z^i be a representation of x^i, then*

$$I(y; z^i) \geq I(y; x^i x^j) - I(x^i; x^j|z^i) - I(x^i; y|x^j) - I(x^j; y|x^i).$$

Proof. Hypothesis (H1): z^i is a representation of x^i: $I(y; z^i|x^i x^j) = 0$.

$$\begin{aligned}
I(y; z^i) &\overset{(P3)}{=} I(y; z^i|x^i x^j) + I(y; x^i x^j; z^i) \\
&\overset{(H1)}{=} I(y; x^i x^j; z^i) \\
&\overset{(P3)}{=} I(y; x^i x^j) - I(y; x^i x^j|z^i) \\
&\overset{(P2)}{=} I(y; x^i x^j) - I(y; x^i|z^i) - I(y; x^j|z^i x^i) \\
&\overset{(P3)}{=} I(y; x^i x^j) - I(y; x^i|z^i) - I(y; x^j|x^i) + I(y; x^j; z^i|x^i) \\
&\overset{(P3)}{=} I(y; x^i x^j) - I(y; x^i|z^i) - I(y; x^j|x^i) + I(y; z^i|x^i) - I(y; z^i|x^i x^j) \\
&\overset{(H1)}{=} I(y; x^i x^j) - I(y; x^i|z^i) - I(y; x^j|x^i) + I(y; z^i|x^i) \\
&\overset{(P1)}{\geq} I(y; x^i x^j) - I(y; x^i|z^i) - I(y; x^j|x^i) \\
&\overset{(Th.B.3)}{\geq} I(y; x^i x^j) - I(x^i; y|x^j) - I(x^i; x^j|z^i) - I(y; x^j|x^i)
\end{aligned}$$

□

Corollary B.1. *Let x^i and x^j be mutually redundant views for y. Let z^i be a representation of x^i that is sufficient for x^j. Then we have:*

$$I(y; z^i) = I(x^i x^j; y).$$

Proof. Hypotheses (H1): z^i and z^j are mutually redundant for y: $I(y; x^i|x^j) + I(y; x^j|x^i) = 0$.
(H2): z^i is sufficient forx^j: $I(x^j; x^i|z^i) = 0$.
Using the Theorem B.4 and Hypotheses (H1)-(H2), we have

$$\begin{aligned}
I(y; z^i) &\overset{(Th.B.4)}{\geq} I(y; x^i x^j) - I(x^i; y|x^j) - I(x^i; x^j|z^i) - I(y; x^j|x^i) \\
&\overset{(H1)}{=} I(y; x^i x^j) - I(x^i; x^j|z^i) \\
&\overset{(H2)}{=} I(y; x^i x^j)
\end{aligned}$$

Since $I(y; z^i) \leq I(y; x^i x^j)$ is a consequence of the data processing inequality, we conclude that $I(y; z^i) = I(y; x^i x^j)$. □

B.3 Consistent Variational Lower Bound for SCMVC

Consistent variational lower bound provides an efficient approximate posterior inference of the latent variable $\vec{z}, y^i$ given an observed value x^i. We maximize the following formulation:

$$\max \log q_\phi(x^i) = D_{KL}(p_\theta(\vec{z}, y^i|x^i)||q_\phi(\vec{z}, y^i|x^i)) + L(\phi, \theta; x^i), \tag{1}$$

as $D_{KL}(p_\theta(\vec{z}, y^i|x^{(i)})||q_{\phi^i}(\vec{z}, y^i|x^i)) \geq 0$, then we have

$$\log q_{\phi^i}(x^i) \geq L(\phi^i, \theta^i; x^i) = \mathbb{E}_{p_{\theta^i}(z,y|x^i)}[-\log p_{\theta^i}(z^i, y^i|x^{(i)}) + \log q_{\phi^i}(x^i, z^i, y^i)],$$

Inspired by [1], which is better at generalization due to its emphasis on disentangled representations and robustness during training, $L(\phi, \theta, x^i)$ can be expressed as follows:

$$\max L_2^i = \mathbb{E}_{p_{\theta^i}(\vec{z},y|x^i)}[\log q_{\phi^i}(x^i|\vec{z}, y^i)] - \gamma D_{KL}(p_{\theta^i}(\vec{z}, y^i|x^i)||q_{\phi^i}(\vec{z}, y^i)), \tag{2}$$

As we assume that

$$q_\phi(x^i|\vec{z}, y^i) = q_\phi(x^i|\vec{z}),\ p_\theta(\vec{z}, y^i|x^i) = p_\theta(y^i|\vec{z})p_\theta(\vec{z}|x^i),$$

$\mathbb{E}_{p_\theta(\vec{z},y^i|x^i)}[\log q_{\phi^i}(x^i|\vec{z}, y^i)]$ is equivalent as

$$\begin{aligned}
\mathbb{E}_{p_\theta(\vec{z},y^i|x^i)}[\log q_{\phi^i}(x^i|\vec{z}, y^i)] &= \iint \sum_y p_\theta(\vec{z}, y^i|x^i) \log q_\phi(x^i|\vec{z}) dz dx \\
&= \iint \sum_y p_\theta(y^i|\vec{z}) p_\theta(\vec{z}|x^i) \log q_\phi(x^i|\vec{z}) dz dx \\
&= \iint p_\theta(\vec{z}|x^i) \log q_\phi(x^i|\vec{z}) dz dx
\end{aligned}$$

For $D_{KL}(p_\theta(z, y^i|x^i||q_\phi(z, y))$,

$$\begin{aligned}
D_{KL}(p_\theta(\vec{z}, y^i|x^i||q_\phi(\vec{z}, y)) &= \sum_y \iint p_\theta(\vec{z}, y^i|x^i) ln \frac{p_\theta(\vec{z}, y^i|x^i)}{q_\phi(\vec{z}, y^i)} dz dx \\
&= \sum_y \iint p_\theta(y^i|\vec{z}) p_\theta(\vec{z}|x^i) ln \frac{p_\theta(y^i|\vec{z}) p_\theta(\vec{z}|x^i)}{q_\phi(\vec{z}|y^i) q_\phi(y^i)} dz dx \\
&= \mathbb{E}_{x\sim\tilde{p}(x)}[\mathbb{E}_{\vec{z}\sim\tilde{p}(\vec{z}|x^i)}[\sum_y p_\theta(y^i|\vec{z}) KL(p_\theta(\vec{z}|x^i)||q_\phi(\vec{z}|y^i)) + KL(p_\theta(y^i|\vec{z})||q_\phi(y^i))]]
\end{aligned}$$

Given

$$q(\vec{z}|x^i) = \frac{1}{\prod_{t=1}^{d} \sqrt{2\pi\sigma_t^{i2}(x^i)}} exp\{-\frac{1}{2}||\frac{z^i - \mu(x^i)}{\sigma(x^i)}||^2\}$$

and

$$q(\vec{z}|y_j^i) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp(-\frac{1}{2}||\vec{z} - \mu_j^i||^2),$$

the objective function of SCMVC can be formulated as follows:

$$\max L_{con}^i \leftrightarrow \max \sum_{m=1}^{n} -|x_m^i - \tilde{x_m^i}|^2 + \sum_{j=1}^{k} \gamma(y_{mj}^i (\sum_{t=1}^{d} \log \sigma_t^2(x)) - \frac{1}{2}||\vec{z} - \mu_j^i||^2 + \log(y_{mj}^i))$$

where d is the dimension of hidden layer features, y_{ij}^v (obtained from z_i^v through softmax layer) is the probability of instance i assigned to class j in view-v, and $\tilde{x}^i$ represents the reconstruction item through the decoder. For μ_j^i, which is initialized randomly, is the mean vector of the class of i-th view.

B.4 Bayes Error Rates for Arbitrary Learned Representations

Theorem B.5. *The Bayes Error Rates for Arbitrary Learned Representations can be minimized by maximizing* $L_c^{ij} on + \beta L_{suf}^{i}$*, formally:*

$$\max L_{con}^{i} + \beta L_{suf}^{ij} \leftrightarrow \min \bar{p}_e. \tag{3}$$

Proof. *Our reparameterization strategy transforms* z^i *into a normal distribution that maximizes* $H_{\theta^*}^{(n)}(z^i)$*. Moreover, we obtain consistent information among views, as expressed in consistent lower bound, which minimizes* $\hat{H}_{\theta^*}^{(n)}(z^i|x^j)$*. As we have* $\hat{I}_{\theta^*}^{(n)}(z^i; x^j) = \hat{H}_{\theta^*}^{(n)}(z^i) - \hat{H}_{\theta^*}^{(n)}(z^i|x^j)$*, the resulting* $\hat{I}_{\theta^*}^{(n)}(z^i; x^j)$ *is maximized. As for minimizing* $I(z^i; x^i|x^j, T)$*, given Proof of the sufficient representation lower bound in Section 3.3 the conclusion can be stated as follows:*

$$\max -D_{KL}(p(z^i|x^i), p(z^j|x^j)) \leftrightarrow \min I(z^i; x^i|x^j, T).$$

Overall, the Bayes Error Rates for Arbitrary Learned Representations are minimized:

$$\max L_{con}^{i} + \beta L_{suf}^{ij} \leftrightarrow \min \bar{p}_e. \tag{4}$$

□

C Sufficient Representation Loss

Based on the assumption that $\{z^i\}_{i=1}^v$ follows a Gaussian distribution, $D_{KL}(p(z^i|x^i), p(z^j|x^j))$ can be directly optimized via minimizing the KL divergence:

$$D_{KL}(p(z^i|x^i), p(z^j|x^j)) = \frac{1}{2}[\mathrm{tr}((\Sigma^i)^{-1}\Sigma^j) + (\mu^i - \mu^j)^{\mathrm{T}}(\Sigma^i)^{-1}(\mu^i - \mu^j) - d + \ln\frac{|\Sigma^i|}{|\Sigma^j|}], \tag{5}$$

where tr stands for the trace of a matrix defined as the sum of the diagonal elements of the matrix, Σ^i, Σ^j are the diagonal matrices where the diagonal elements are the variances of z^i and z^j in each dimension respectively, and d is the dimension of embeddings $\{z^i\}_{i=1}^v$. The mutual information between the two representations can be maximized by using any sample-based differentiable mutual information lower bound like InfoNCE. In our methods we found that the predictive information term didn't play a role in improving final clustering results. As we focus on discarding superfluous information to finetune the model, we discard the signal-relevant term during fintuning. It's interesting to further explore this discovery in the future work.

D More Details of the Proposed Methods

D.1 Implementation Details

We use the convolutional autoencoder (CAE) for each view to learn embedded features. The encoder is Input-Conv_4^{32}-Conv_4^{64}-Conv_4^{64}-Fc_{200} and the decoders are symmetric with the corresponding encoders. For the comparing methods, we use open-source codes with the settings recommended by the authors. For all datasets an encoder and a decoder are shared for all views. The experiments were conducted utilizing a Windows PC installed with an Intel(R) Core(TM) i5-12600K CPU@3.69 GHz, 32.0 GB RAM, and GeForce RTX 3070ti GPU (8 GB cache). All baselines were tuned to their highest performance according to the respective papers to ensure fair comparison. We set β and γ to 0.1 for all experiments. To better understand our approach, we provide the framework of SCMVC and SUMVC in Figs. 1-2.

D.2 Visualization of Learned Generative Models

We show the reconstructed pictures and sampled pictures generated on the Multi-Mnist and Multi-COIL-20 respectively (Figs. 3-5). We utilize the trained model, set the batch size to 32, and display the images of each view for three epochs.

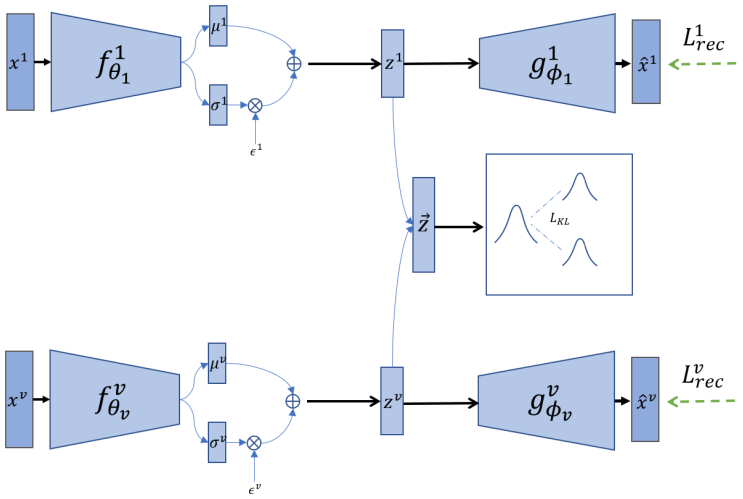


Figure 1: The framework of SCMVC. For the v-th view, x^v denotes the input data, z^v denotes the embedded features, $f^v_{\theta_v}$ and $g^v_{\phi_v}$ are the encoder and the decoder of v-th view respectively.

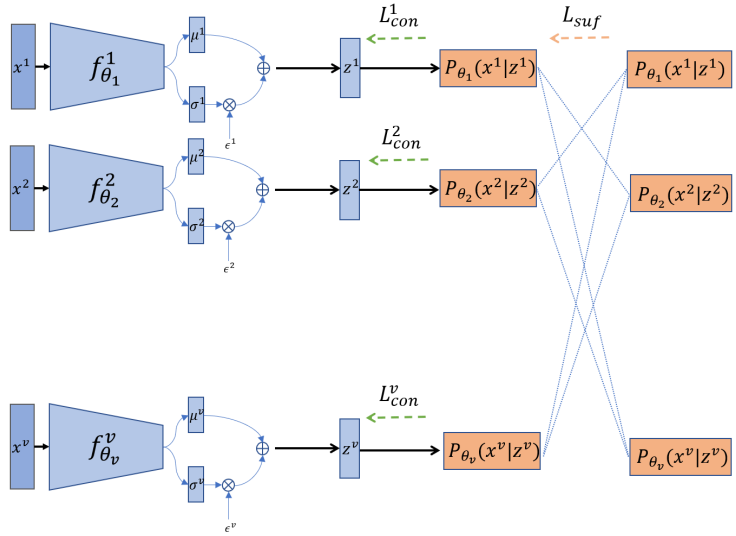
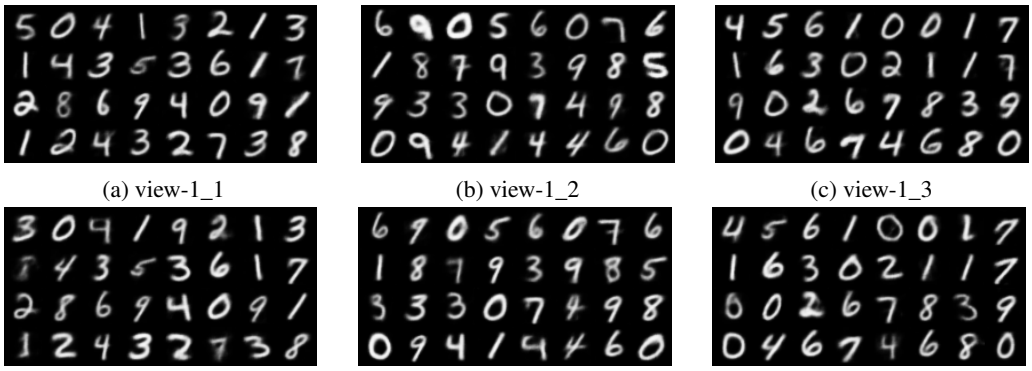


Figure 2: The framework of SUMVC. We utilize a collaborative training approach whereby each view sequentially serves as a guide to facilitate other views' learning process.

D.3 Limitation of the Proposed Methods

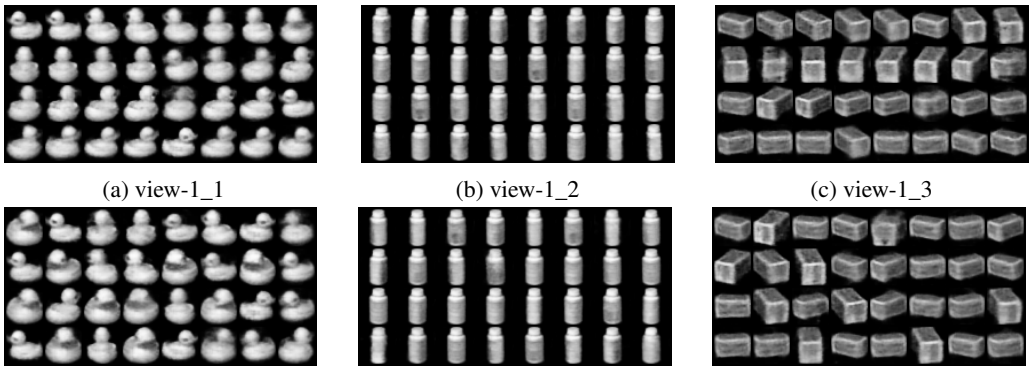
While our proposed methods have demonstrated promising results in addressing the multi-view clustering research problem and have provided valuable insights into strengthening consistency and decreasing view redundancy, it is important to acknowledge several limitations that should be considered. In particular, our model does not perform well on datasets with particularly strong heterogeneity between views, such as huge differences in dimensions of different views. This is because variational inference assumes that all samples in the data come from a common probability distribution. However, this assumption may not hold true when the data is highly heterogeneous, and the distribution may exhibit a range of subgroups and variations. In such cases, the use of variational inference may not be accurate or effective. To the best of our knowledge, this issue remains unexplored within the existing literatures. In the future, we will consider employing ensemble variational inference or gaussian process mixture model as a potential solution to address this problem and achieve a more robust model.



(a) view-1_1 (b) view-1_2 (c) view-1_3

(d) view-2_1 (e) view-2_2 (f) view-2_3

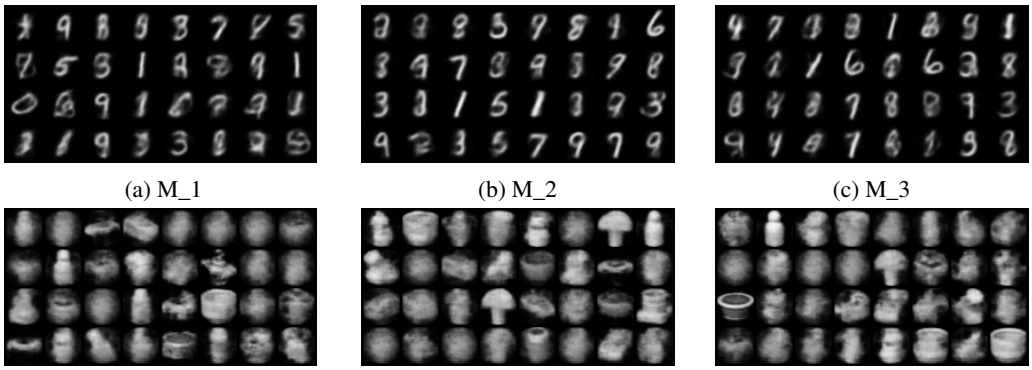
Figure 3: The reconstructed pictures generated on Multi-Mnist.



(a) view-1_1 (b) view-1_2 (c) view-1_3

(d) view-2_1 (e) view-2_2 (f) view-2_3

Figure 4: The reconstructed pictures generated on Multi-COIL-20.



(a) M_1 (b) M_2 (c) M_3

(d) C_1 (e) C_2 (f) C_3

Figure 5: The sampled pictures generated on the Multi-Mnist and Multi-COIL-20.

References

[1] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[2] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2020.