Contrastive Max-correlation for Multi-view Clustering

Yanghao Deng, Zenghui Wang, and Songlin Du(✉)

School of Automation, Southeast University, Nanjing 210096, China
{yhdeng0301, scholarzhwang}@163.com, sdu@seu.edu.cn

Abstract. Multi-view clustering exhibits advantages over single-view clustering due to its ability to fully utilize complementary information between multiple views. However, most mainstream methods have the following two drawbacks: 1) Ignoring structural conflicts between views leads to a deterioration in clustering performance, because merging a certain view actually worsens the clustering results; 2) Rather than globally extracting the maximum correlation between views, their approaches center on individual instances, consequently making models more susceptible to interference from local noise points. To address these issues, this paper proposes a novel framework, entitled Contrastive Max-correlation for Multi-view Clustering (CMMC) for robust multi-view clustering. In particular, the network framework incorporates two effective methods. The first method, maximum structure correlation learning, enhances the downstream task representations by incorporating complementary structural information. Additionally, the framework achieves simultaneous mining of view correlations and alignment of views through the global max-correlation contrastive learning method. As the above methods operate globally, CMMC can effectively reduce the impact of noise information. Experiments on various types of multi-view datasets demonstrate that CMMC outperforms existing methods in terms of clustering accuracy and robustness.

Keywords: Deep multi-view clustering · Contrastive learning · Canonical correlation analysis · Unsupervised learning.

1 Introduction

In recent years, the significance of multi-view clustering has been steadily growing. In modern society, data gathered for real-world applications typically originates from diverse domains, sensors, or feature extractors [12]. Multi-view clustering excels precisely because it can extract and leverage the complementary and consistent information inherent within such multi-source data, thereby demonstrating superior performance in downstream tasks. Additionally, multi-view clustering falls under the category of unsupervised learning, offering the notable advantage of functioning without the need for labeling information during the training process. This circumvents the challenge of acquiring costly labeled data, ultimately leading to significant cost savings.

Traditional multi-view clustering methods, based on how they combine the multi-view information [3], can be categorized into five classes: 1) common eigenvector matrix, 2) common coefficient matrix, 3) common indicator matrix, 4) direct view combination, 5) view combination after projection. Compared with these traditional shallow methods, the deep neural network is a more promising approach because of its excellent nonlinear mapping capability and its flexibility in different scenarios [20]. In order to fully utilize the complementary and consistent information of multi-view data, many prominent methods have been proposed. Although these methods have made some progress in multi-view tasks, the following open problems still exist: 1) Many methods [27,28,30] focus too much on aligning or not aligning local points when introducing contrastive learning methods for alignment task, which often weakens the robustness of models due to the presence of individual noisy data points; 2) Some methods [12,33], in their pursuit of achieving consistent representation, overlook structural conflicts among different viewpoints. They forcefully merge these viewpoints, thereby compromising the complementary information inherent in diverse views. Therefore, after integrating a particular view, there may be a decrease in model performance, which contradicts the concept of multi-view clustering where more views lead to more information and higher performance.

Based on the above observations, we propose a novel multi-view representation learning framework for clustering to address the aforementioned issues, entitled Contrastive Max-correlation for Multi-view Clustering (CMMC). Overall, CMMC aims to globally align views and maximize structural complementary information across views. In detail, we first introduce the variational autoencoder to obtain the view-specific representations by reconstructing the origin data. Then, the novel method, entitled maximum structure correlation learning, is proposed to reshape new representations with structural information and utilize the reshaped representation to guide the representation of other views to learn structural complementary information of the view. Furthermore, we introduce Deep Canonical Correlation Analysis (DCCA) [2] into contrastive learning for the first time to align multi-view data and maximize correlation across views globally. Compared with previous work, our contributions are listed as follows:

- We design a flexible framework, entitled Contrastive Max-correlation for Multi-view Clustering (CMMC), which is able to maximize structural complementary information across views.
- As far as we know, global max-correlation contrastive learning could be the first attempts which introduce DCCA into contrastive learning to build a novel contrastive loss with the ability to align multi-view data and maximize correlation across views simultaneously.
- All of our operations are from a global perspective. As a result, the influence of noise points is effectively reduced and the robustness of the model is improved. Experiments show that the proposed method outperforms several state-of-the-art methods on five public datasets.

2 Related Works

2.1 Deep Multi-view Clustering

Although traditional multi-view clustering algorithms have achieved promising results in some tasks, multi-view clustering algorithms based on deep learning have emerged as the mainstream research direction due to their superior representation learning ability. Similar to traditional methods, deep multi-view clustering also mines complementarity and consistency information between multiple views to enhance model performance. For example, the algorithms proposed by [16] and [31] learn a consistent representation by mining mutual information between multi-view representations. Conversely, [4, 10, 11] exploit the deep spectral clustering algorithm proposed by [21] to mine consistency information of non-convex and more complex datasets. However, many existing deep multi-view clustering algorithms struggle to handle datasets with more than two views, and even exhibit a phenomenon where the results deteriorate as the number of dataset views increases [28]. This may be attributed to the conflicting structures among views during the process of mining more view complementarity and consistency information, leading to a dip in model performance.

In contrast to the above methods, our algorithm CMMC neutralizes structural conflicts between two views and establishes structural robustness among multiple views. To achieve this goal, CMMC first reconstructs the representation of a view using its structural information, and then utilize the reconstructed representation to guide the learning of other views' representations. In **Table 2**, we verify that this approach can effectively eliminate the phenomenon of conflicting view structures.

2.2 Contrastive Learning

Contrastive learning has recently made significant strides in self-supervised representation learning. These methods rely heavily on numerous distinct pairwise representation comparisons. Specifically, they aim to maximize the similarities among positive pairs while simultaneously minimizing those among negative pairs in a latent feature space. [32] combined information theory with contrastive learning to alleviate cross-view discrepancies and learn consensus semantics. [19] employed intra- and inter-view contrastive learning in different instead of the same space, thus being in favor of the intra-view information and cross-view consistency. [15] proposed a framework for contrastive learning at both the cluster-level and instance-level, respectively conducted in the row and column space, by maximizing the similarities of positive pairs while minimizing those of negative ones. [5] developed a contrastive learning framework solely at the cluster-level, but did so by identifying a superior surrogate for the source data of positive and negative sample pairs, ensuring that cluster allocation pairs in the same view are drawn together while pushing cluster allocation pairs into another view.

Despite these methods' adept utilization of the benefits of contrastive learning, they overlook the impact of view-specific local noise when constructing positive and negative sample pairs. These local noises disrupt model training, causing

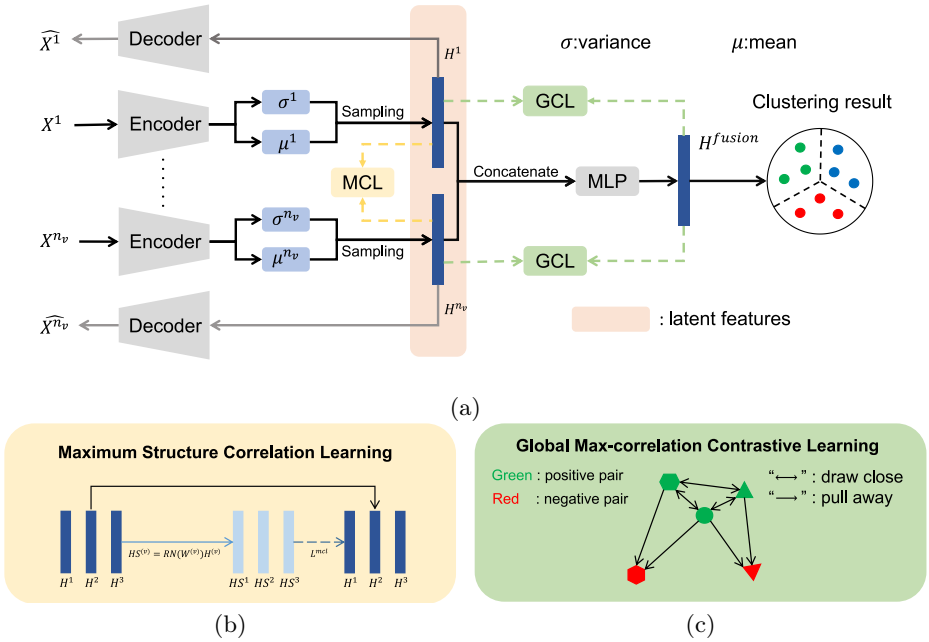


Fig. 1: (a) The framework of CMMC. We employ a set of encoder networks to map the input data $\mathbf{X}^{(v)}$ into the mean and variance within the latent space. Reparameterization trick is employed to sample a latent feature $\mathbf{H}^{(v)}$. Then, these latent features are fused through the MCL and GCL methods to generate $\mathbf{H}^{(fusion)}$, which contains consistent and complementary information from each view. Finally, the learned $\mathbf{H}^{(fusion)}$ is utilized to generate the clustering outcomes for k-means clustering. **(b) Maximum Structure Correlation Learning (MCL)** aims to learn complementary structural information. **(c) Global Max-Correlation Contrastive Learning (GCL)** aligns multi-view data and maximize correlation across views globally. Different shapes represent different views.

it to veer off course and learn incorrect patterns or features, ultimately diminishing the model's accuracy on test data. Therefore, we propose integrating DCCA into contrastive learning to globally construct positive and negative sample pairs to mitigate the effects of local noise.

3 Method

Let $\mathbf{X} = \left\{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^{(n_v)}\right\}$ be a multi-view dataset with n_v different views. For the v-th view, $\mathbf{X}^{(v)} \in \mathbb{R}^{n \times d_v} (1 \leq v \leq n_v)$, where n is the number of data points and d_v denotes the dimension. The latent representations obtained from Variational Auto-Encoding (VAE), corresponding negative pairs representations,

and fusion representation are denoted by $\left\{\mathbf{H}^{(v)}, \overline{\mathbf{H}}^{(v)}, \mathbf{H}^{(fusion)}\right\} \in \mathbb{R}^{n\times k}$ respectively. The whole framework of CMMC is shown in Figure 1(a).

3.1 Variational Representation Learning

In multi-view clustering tasks, autoencoders have been widely employed to extract representations from raw features. In our study, we introduce the VAE to enhance robustness of model when confronted with noise and variations.

The VAE consists primarily of two processes: encoding $q_\phi(\mathbf{H} \mid \mathbf{X})$ and decoding $\mathrm{p}_\theta(\mathbf{X} \mid \mathbf{H})$. Generally, assuming the data follows a Gaussian distribution, the encoder network $q_\phi(\mathbf{H} \mid \mathbf{X})$ takes the original data as input and produces two outputs, μ and σ, representing the mean and variance of the Gaussian distribution, respectively. Then, using the reparameterization trick, a latent variable $\mathbf{H}$ is generated. Finally, the latent variable $\mathbf{X}$ is used by the decoder network $\mathrm{p}_\theta(\mathbf{X} \mid \mathbf{H})$ to reconstruct the original data, where ϕ and θ represent the parameters of the encoder and decoder networks, respectively.

The aim of VAE is to obtain the true posterior probability distribution $\mathrm{p}_\theta(\mathbf{H} \mid \mathbf{X})$ by performing an approximate estimation through finding the distribution $q_\phi(\mathbf{H} \mid \mathbf{X})$ that is closest to it, which can be realized by optimizing the constraint parameters θ and ϕ [13]:

$$\log p_\theta(\mathbf{X}) = D_{\mathrm{KL}}\left(q_\phi(\mathbf{H} \mid \mathbf{X}) \| p_\theta(\mathbf{H} \mid \mathbf{X})\right) + \mathcal{L}^{ELBO}(\theta, \phi; \mathbf{X}), \tag{1}$$

where $\log p_\theta(\mathbf{X})$ is the logarithm of the probability of occurrence of the sample data. The second term is called Evidence Lower Bound (ELBO). In the context where the constant term on the left-hand side remains unchanged, maximizing the ELBO minimizes the Kullback-Leibler (KL) divergence term. The final objective function of a VAE aims to minimize the sum of the reconstruction loss and the KL divergence. After transformation and derivation, it can be expressed as:

$$\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{H}|\mathbf{X})}\left[\log p_\theta(\mathbf{X} \mid \mathbf{H})\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{H} \mid \mathbf{X}) \| p_\theta(\mathbf{H})\right). \tag{2}$$

As shown in Figure 1(a), there are a total of n_v VAEs, so the basic VAE loss of our model is the sum of the VAE losses of the n_v views

$$\begin{aligned} \mathcal{L}^{vae} = \sum_{i=1}^{n_{\mathrm{v}}} \Big(& \mathbf{E}_{q_\phi\left(\mathbf{H}^{(i)}|\mathbf{X}^{(i)}\right)}\left[\log p_\theta\left(\mathbf{X}^{(i)} \mid \mathbf{H}^{(i)}\right)\right] - \\ & \alpha\, D_{\mathrm{KL}}\left(q_\phi\left(\mathbf{H}^{(i)} \mid \mathbf{X}^{(i)}\right) \| p_\theta\left(\mathbf{H}^{(i)}\right)\right)\Big), \end{aligned} \tag{3}$$

where α is a trade-off parameter, and is fixed to 5e-7.

3.2 Maximum Structure Correlation Learning

The structural information across different views is a crucial aspect to consider for ensuring complementarity and consistency in multi-view learning. When integrating multi-view information, significant disparities in structural information

among views can lead to conflicts and subsequent performance degradation. To address this challenge, we propose the Maximum Structure Correlation Learning (MCL) method. The detailed implementation of the method is illustrated in Figure 1(b).

After the latent presentations $\mathbf{H}^{(v)}$ extracted by VAEs, our goal is to enhance the representations of different views by incorporating complementary structural information from other views. A new embedding space is reshaped and utilized to guide other views to learn the structural complementary information from the view. In details, we apply the k-NN graph to construct the non-negative affinity matrix $\mathbf{W}^{(v)} \in R^{n \times n}$ as the structural information. The matrix is defined as

$$\mathbf{W}_{ij}^{v} = \begin{cases} \exp\left(-\frac{\left\|h_i^v - h_j^v\right\|_2^2}{2\sigma^2}\right), & \text{if } h_i^v, h_j^v \text{ are connected.} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

To be exact, the similarity between h_i^v and h_j^v is computed by a Gaussian kernel with a scale $\sigma > 0$ if h_j^v falls into the k neighborhood of h_i^v and the selection of k is fixed at 40. Using the similarity matrix, we can reshape the new representation $\mathbf{HS}^{(v)}$ with structural information through a combination of its k nearest neighbors:

$$\mathbf{HS}^{(v)} = \mathbf{RN}\left(\mathbf{W}^{(v)}\right)\mathbf{H}^{(v)}. \tag{5}$$

$\mathbf{RN}$ means that the matrix $\mathbf{W}^{(v)}$ is normalized by rows, i.e., the elements of each row of the matrix $\mathbf{W}^{(v)}$ are divided by the sum of that row.

Ultimately, we use cross-entropy to guide the representations of other views to learn complementary structural information from the current view, represented as $D_{\mathrm{KL}}\left(\mathbf{HS}^{(i)} \| \mathbf{H}^{j}\right)$. Additionally, cross-entropy is employed to effectively reduce the variability of the structural space between different views, which is represented mathematically as $D_{\mathrm{KL}}\left(\mathbf{W}^{(i)} \| \mathbf{W}^{(j)}\right)$, thereby ensuring a more consistent and coherent representation across views. During the experiment, we found that the loss function changes too dramatically, and using an exponential function as the basis for MCL can increase the smoothness of the loss function, which is beneficial for the stability and smoothness of the algorithm. Therefore, the $\mathcal{L}^{mcl}$ objective is defined by

$$\mathcal{L}^{mcl} = exp\Big(-\sum_{i=1}^{n_v}\sum_{j=1}^{n_v}\mathbf{HS}^{(i)}log\mathbf{H}^{(j)} - \sum_{i=1}^{n_v-1}\sum_{j=i+1}^{n_v}\mathbf{W}^{(i)}log\mathbf{W}^{(j)}\Big). \tag{6}$$

3.3 Global Max-correlation Contrastive Learning

In order to explore the maximum correlation across views as well as reduce noise information globally, we introduce DCCA into contrastive learning and revised the selection method for positive and negative samples to create a new unsupervised loss function. This novel loss function ignores the influence of noisy

samples at the instance-level and maintains consistency across views. The detailed implementation of the method is illustrated in Figure 1 (c).

According to [2], the correlation of pairs, $\operatorname{corr}\left(\mathbf{H}^{(1)}, \mathbf{H}^{(2)}\right)$, is computed by

$$\begin{aligned} &\operatorname{corr}\left(\mathbf{H}^{(1)}, \mathbf{H}^{(2)}\right) = \|T\|_{\mathrm{tr}} = \operatorname{tr}\left(T'T\right)^{1/2} \\ &T = \hat{\Sigma}_{11}^{-1/2} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2}, \end{aligned} \tag{7}$$

where $\left(\hat{\Sigma}_{11}, \hat{\Sigma}_{22}\right)$ represents covariance and $\hat{\Sigma}_{12}$ is cross-covariance.

This method for measuring data correlation can better handle nonlinear and complex data. Moreover, by considering a batch of data from different views of the dataset rather than individual sample points from each view, it effectively mitigates the impact of local noise points and enhances the maximum correlation between views. Therefore,

$$d\left(\mathbf{H}^{(1)}, \mathbf{H}^{(2)}\right) = \frac{1}{\operatorname{corr}\left(\mathbf{H}^{(1)}, \mathbf{H}^{(2)}\right)} \tag{8}$$

can be regarded as an innovative distance measure, capable of replacing the Euclidean and Cosine distances. The validity experiments are presented in **Table 4**.

We take the paired view data as positive samples, and the randomly selected samples from the corresponding view and fused view, respectively, as negative samples. As a result, the positive loss function and negative loss function are

$$\mathcal{L}^{pos} = \sum_{i=1}^{n_v-1} \sum_{j=i+1}^{n_v} d^2\left(\mathbf{H}^{(i)}, \mathbf{H}^{(j)}\right) + \sum_{i=1}^{n_v} d^2\left(\mathbf{H}^{(i)}, \mathbf{H}^{(fusion)}\right) \tag{9}$$

and

$$\mathcal{L}^{neg} = \sum_{i=1}^{n_v} \max\left(m - d\left(\mathbf{H}^{(i)}, \overline{\mathbf{H}}^{(i)}\right), 0\right)^2 + \max\left(m - d\left(\mathbf{H}^{(fusion)}, \overline{\mathbf{H}}^{(i)}\right), 0\right)^2, \tag{10}$$

where m is a margin which enforces the distance of negatives to be moderately large. To avoid laboriously parameter selection, we suggest automatically determining the appropriate m for each dataset during its initial stages [29]. This process can be expressed mathematically as

$$m = \frac{1}{N_p} \sum_{i=1}^{n_v-1} \sum_{j=i+1}^{n_v} d\left(\mathbf{H}^{(i)}, \mathbf{H}^{(j)}\right) + \frac{1}{N_n} \sum_{i=1}^{n_v} d\left(\mathbf{H}^{(i)}, \overline{\mathbf{H}}^{(i)}\right), \tag{11}$$

where N_p and N_n denote the number of positive and negative pairs, respectively. By Eq. 11, we can calculate $N_p = n_v(n_v - 1)/2$ and $N_n = n_v$.

Consequently, our max-correlation contrastive learning loss function is

$$\mathcal{L}^{gcl} = \frac{1}{2N}\left(\mathcal{L}^{neg} + \beta\mathcal{L}^{pos}\right), \tag{12}$$

where $N = N_p + N_n$ denotes the number of contrastive pairs and β is a pre-defined hyper-parameter whose values we fix at 130.

Overall, the loss of our method consists of the above three parts:

$$\mathcal{L} = \mathcal{L}^{mcl} + \mathcal{L}^{gcl} + \mathcal{L}^{vae}. \tag{13}$$

4 Experiments

4.1 Implementation Details

We implement our model using PyTorch 1.12.1 and conduct all evaluations on an NVIDIA 3070 GPU. The Adam optimizer is employed with an initial learning rate of 0.001, without a scheduler or weight decay. For the experiments presented in **Table 1**, we train the model for 150 epochs across all datasets, with a batch size of 1024. For the experiments in **Table 2**, we pre-train the models for 200 epochs to minimize the reconstruction loss and then fine-tune the model for 100 epochs using mini-batches of size 256. The evaluation metrics include four measures: accuracy (ACC), normalized mutual information (NMI), adjusted Rand index (ARI), and purity (PUR).

4.2 Datasets

Four widely-used multi-view datasets, including three handcraft-feature-based (Scene15 [8], Reuters [1], Caltech101 [7]) and the NoisyMNIST [25] datasets, are chosen to evaluate the effectiveness of CMMC. One five-view dataset (Caltech [6]) is used to test whether CMMC can fully utilize the complementary information of multiple views to improve the performance of model in the face of increased views.

4.3 Comparison Methods

For these four two-view datasets, we compare CMMC with 6 multi-view clustering baselines including CCA [24], DCCA [2], DAIMC [9], EERIMVC [18], SURE [29], and ProImp [14]. For the five-view dataset, the comparison methods include 6 state-of-the-art methods, including EAMC [33], CDIMC-net [26], COMPLETER [17], SiMVC [23], CoMVC [23], MFLVC [27], and GCFAgg [28].

4.4 Result Analysis

Table 1 presents the clustering results of four widely-used multi-view datasets which are used to evaluate the effectiveness of CMMC. As shown in this table,

Table 1: Multi-view clustering comparisons on four widely-used multi-view datasets. The best result in each row is shown in **bold** and the second-best is underlined. The average clustering performance is reported.

Evaluation metrics	Datasets	ACC	NMI	ARI
CCA	Scene-15	36.37	36.91	19.82
	Caltech-101	20.25	45.41	16.34
	Reuters	44.31	20.34	14.52
	NoisyMNIST	71.31	52.60	48.46
DCCA	Scene-15	36.61	39.20	21.03
	Caltech-101	27.60	47.84	30.86
	Reuters	47.95	26.57	12.71
	NoisyMNIST	89.64	88.33	83.95
DAIMC	Scene-15	32.09	33.55	17.42
	Caltech-101	26.40	49.18	19.00
	Reuters	40.78	21.15	15.98
	NoisyMNIST	38.40	34.66	22.98
EERIMVC	Scene-15	39.60	38.99	22.06
	Caltech-101	23.98	45.61	17.19
	Reuters	33.21	14.28	3.9
	NoisyMNIST	65.66	57.60	51.34
SURE	Scene-15	40.95	43.19	25.01
	Caltech-101	34.59	48.30	48.79
	Reuters	49.06	29.91	23.56
	NoisyMNIST	98.36	95.38	96.43
ProImp	Scene-15	43.61	45.02	26.84
	Caltech-101	–	–	–
	Reuters	56.54	39.35	32.77
	NoisyMNIST	**99.17**	**97.48**	**98.18**
CMMC(Ours)	Scene-15	**47.33**	**46.05**	**29.42**
	Caltech-101	**36.33**	**51.84**	**54.38**
	Reuters	**60.57**	**41.89**	**35.42**
	NoisyMNIST	99.05	97.12	97.92

although our method is slightly inferior to the ProImp on the NoisyMNIST dataset, we get the best results on all other datasets. Especially, CMMC achieves an ACC improvement of **8.53**% and **7.13**% on the Scene and Reuters dataset, respectively. **Table 2** presents the clustering results on Caltech with different views. Overall, our proposed CMMC outperforms all the competitors on all the metrics and datasets. In addition, unlike GCFAgg, it can be observed that our method improves all the metrics as the views increases. These results illustrate that our method can fully utilize the complementary information of multiple views to improve the performance of model in the face of increased views.

4.5 Ablation Study

According to the overall multi-view clustering loss in Eq. 13, three distinct loss components are involved. To ascertain the significance of each component in CMMC, we conducted ablation studies under identical experimental conditions to isolate the necessity of each component. As depicted in **Table 3**, we observe

Table 2: Results of all methods on Caltech with different views. "-XV" represents that there are X views.

Evaluation metrics	Datasets	ACC	NMI	PUR
EAMC	Caltech-2V	0.490	0.398	0.540
	Caltech-3V	0.558	0.445	0.576
	Caltech-4V	0.687	0.610	0.719
	Caltech-5V	0.760	0.691	0.785
CDIMC-net	Caltech-2V	0.419	0.256	0.427
	Caltech-3V	0.389	0.214	0.398
	Caltech-4V	0.356	0.205	0.370
	Caltech-5V	0.318	0.173	0.342
COMPLETER	Caltech-2V	0.437	0.391	0.552
	Caltech-3V	0.489	0.446	0.594
	Caltech-4V	0.391	0.355	0.516
	Caltech-5V	0.431	0.431	0.597
SiMVC	Caltech-2V	0.508	0.471	0.557
	Caltech-3V	0.569	0.495	0.591
	Caltech-4V	0.619	0.536	0.630
	Caltech-5V	0.719	0.677	0.729
CoMVC	Caltech-2V	0.466	0.426	0.527
	Caltech-3V	0.541	0.504	0.584
	Caltech-4V	0.568	0.569	0.646
	Caltech-5V	0.700	0.687	0.746
MFLVC	Caltech-2V	0.606	0.528	0.616
	Caltech-3V	0.631	0.566	0.639
	Caltech-4V	0.733	0.652	0.734
	Caltech-5V	0.804	0.703	0.804
GCFAgg	Caltech-2V	0.664	0.501	0.664
	Caltech-3V	0.640	0.535	0.653
	Caltech-4V	0.734	0.661	0.734
	Caltech-5V	0.834	0.733	0.834
CMMC(Ours)	Caltech-2V	**0.666**	**0.538**	**0.666**
	Caltech-3V	**0.724**	**0.590**	**0.724**
	Caltech-4V	**0.776**	**0.688**	**0.776**
	Caltech-5V	**0.876**	**0.774**	**0.876**

that the variational autoencoder loss $\mathcal{L}^{vae}$ alone produces only rudimentary results. Both $\mathcal{L}^{mcl}$ and $\mathcal{L}^{gcl}$ lead to corresponding improvements, but the optimal result is achieved only when $\mathcal{L}^{vae}$, $\mathcal{L}^{mcl}$, and $\mathcal{L}^{gcl}$ are combined.

To assess the effectiveness of Eq. 8 as a distance metric, we conducted an additional experiment on Scene-15 dataset for validation purposes. The result is depicted in **Table 4**. In this experiment, we maintained all other conditions constant while varying the distance metric used to calculate distances, including Cosine distance, Euclidean distance, and our proposed Eq. 8. The experimental results indicate that the performance using Cosine distance was the poorest, followed by Euclidean distance, while our proposed Eq. 8 demonstrated the best performance. This finding suggests that our distance metric is more adept at handling complex datasets, thereby achieving superior performance by mitigating the influence of local noise.

Table 3: Ablation study of three losses on Scene-15.

$\mathcal{L}^{vae}$	$\mathcal{L}^{mcl}$	$\mathcal{L}^{gcl}$	ACC	NMI	ARI
✓			25.74	21.90	10.95
✓	✓		27.49	23.39	11.35
✓		✓	40.03	41.45	24.13
✓	✓	✓	**47.33**	**46.05**	**29.42**

Table 4: Ablation study of three distance measures on Scene-15.

	ACC	NMI	ARI
Cosine	19.33	15.54	4.89
Euclidean	45.04	45.18	27.47
Corr(Ours)	**47.33**	**46.05**	**29.42**

4.6 Parameter Sensitivity Analysis

As discussed above, our method incorporates two balancing parameters: the variational representation learning trade-off parameter α and the contrastive learning trade-off parameter β. While CMMC with fixed parameter values has demonstrated promising performance, it is crucial to investigate the influence of these parameters and fully realize the potential of our method. As shown in **Figure 2**, in this experiment, we vary α within the range of {5e-1, 5e-3, 5e-5, 5e-7, 5e-9} and β in the range of {100, 110, 120, 130}. Notably, we observe that our method exhibits insensitivity to clustering results in a certain parameter range. Specifically, we find that the maximum value is attained when $\alpha = 5e - 7$ and $\beta = 130$, indicating the importance of fine-tuning these parameters for optimal performance.

4.7 Visualizations

To validate the effectiveness of our latent feature $\mathbf{H}^{fusion}$ in integrating the maximum correlation between views, we conduct t-SNE visualization on features learned by CMMC on the NoisyMNIST dataset after convergence. As shown in **Figure 3**, although the view-specific representation $\mathbf{H}$ can form relatively clear cluster boundaries, some examples cannot be classified into the correct clusters. In contrast, our $\mathbf{H}^{fusion}$ can significantly alleviate this issue, with clearer cluster boundaries.

5 Conclusion

In this paper, we introduce a novel framework, Contrastive Max-correlation for Multi-view Clustering (CMMC), which addresses the challenges of multi-view clustering, particularly the structural conflicts and local noise interference often

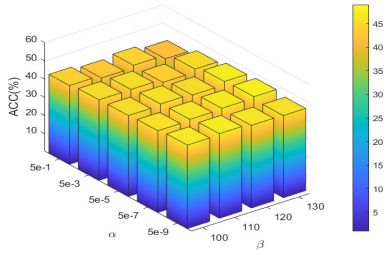
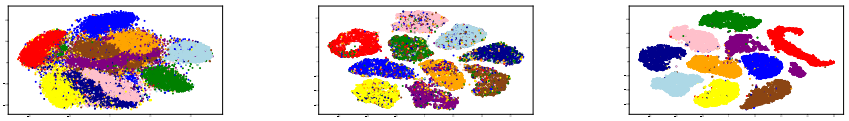


Fig. 2: The effects of α and β on Scene-15.



(a) X **(b) H** **(c) $\mathbf{H}^{fusion}$**

Fig. 3: The t-SNE visualization results on the NoisyMNIST dataset of different feature representations on different layers after convergence. (a) Illustrates the NoisyMNIST dataset in its original state. (b) Represents the view-specific representation $\mathbf{H}$ refined by VAEs. (c) Depicts the learned $\mathbf{H}^{fusion}$ used for downstream tasks.

encountered in existing methods. Our framework incorporates two key components: Maximum Structure Correlation Learning (MCL) and Global Max-correlation Contrastive Learning (GCL). MCL enhances representations by incorporating complementary structural information from other views, thereby reducing conflicts and improving clustering performance. GCL introduces Deep Canonical Correlation Analysis (DCCA) into contrastive learning to globally align views and reduce noise information, enhancing robustness and performance. Experiments on various multi-view datasets demonstrate the superiority of CMMC over existing methods.

References

1. Amini, M.R., Usunier, N., Goutte, C.: Learning from multiple partially observed views-an application to multilingual text categorization. Advances in Neural Information Processing Systems pp. 28–36 (2009)
2. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: Proceedings of the International Conference on Machine Learning. pp. 1247–1255 (2013)
3. Chao, G., Sun, S., Bi, J.: A survey on multi-view clustering. arXiv preprint arXiv:1712.06246 (2017)
4. Chen, J., Mao, H., Peng, D., Zhang, C., Peng, X.: Multiview clustering by consensus spectral rotation fusion. IEEE Transactions on Image Processing **32**, 5153–5166 (2023)

5. Chen, J., Mao, H., Woo, W.L., Peng, X.: Deep multiview clustering by contrasting cluster assignments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16752–16761 (2023)
6. Fei-Fei, L., Fergus, R., Perona, P.: 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. pp. 178–178 (2004)
7. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(4), 594–611 (2006)
8. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 524–531 (2005)
9. Hu, M., Chen, S.: Doubly aligned incomplete multi-view clustering. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 2262–2268 (2018)
10. Huang, Z., Zhou, J.T., Peng, X., Zhang, C., Zhu, H., Lv, J.: Multi-view spectral clustering network. In: Proceedings of the International Conference on International Joint Conferences on Artificial Intelligence. p. 2563–2569 (2019)
11. Huang, Z., Zhou, J.T., Zhu, H., Zhang, C., Lv, J., Peng, X.: Deep spectral representation learning from multi-view data. IEEE Transactions on Image Processing **30**, 5352–5362 (2021)
12. Jin, J., Wang, S., Dong, Z., Liu, X., Zhu, E.: Deep incomplete multi-view clustering with cross-view partial sample and prototype alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11600–11609 (2023)
13. Kingma, D.P.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
14. Li, H., Li, Y., Yang, M., Hu, P., Peng, D., Peng, X.: Incomplete multi-view clustering via prototype-based imputation. arXiv preprint arXiv:2301.11045 (2023)
15. Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J.T., Peng, X.: Contrastive clustering. In: Proceedings of the AAAI conference on artificial intelligence. pp. 8547–8555 (2021)
16. Lin, Y., Gou, Y., Liu, X., Bai, J., Lv, J., Peng, X.: Dual contrastive prediction for incomplete multi-view representation learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4447–4461 (2022)
17. Lin, Y., Gou, Y., Liu, Z., Li, B., Lv, J., Peng, X.: Completer: Incomplete multi-view clustering via contrastive prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11174–11183 (2021)
18. Liu, X., Li, M., Tang, C., Xia, J., Xiong, J., Liu, L., Kloft, M., Zhu, E.: Efficient and effective regularized incomplete multi-view clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(8), 2634–2646 (2020)
19. Lu, Y., Lin, Y., Yang, M., Peng, D., Hu, P., Peng, X.: Decoupled contrastive multi-view clustering with high-order random walks. In: Proceedings of the AAAI conference on artificial intelligence. pp. 14193–14201 (2024)
20. Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X., Philip, S.Y., He, L.: Deep clustering: A comprehensive survey. IEEE Transactions on Neural Networks and Learning Systems (2024)
21. Shaham, U., Stanton, K., Li, H., Nadler, B., Basri, R., Kluger, Y.: Spectralnet: Spectral clustering using deep neural networks. arXiv preprint arXiv:1801.01587 (2018)
22. Tang, H., Liu, Y.: Deep safe incomplete multi-view clustering: Theorem and algorithm. In: International Conference on Machine Learning. pp. 21090–21110 (2022)

23. Trosten, D.J., Lokse, S., Jenssen, R., Kampffmeyer, M.: Reconsidering representation alignment for multi-view clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1255–1265 (2021)
24. Vinokourov, A., Cristianini, N., Shawe-Taylor, J.: Inferring a semantic representation of text via cross-language correlation analysis. Advances in Neural Information Processing Systems p. 1497–1504 (2002)
25. Wang, W., Arora, R., Livescu, K., Bilmes, J.: On deep multi-view representation learning. In: Proceedings of the International Conference on Machine Learning. pp. 1083–1092 (2015)
26. Wen, J., Zhang, Z., Xu, Y., Zhang, B., Fei, L., Xie, G.S.: Cdimc-net: Cognitive deep incomplete multi-view clustering network. In: Proceedings of the International Conference on International Joint Conferences on Artificial Intelligence. pp. 3230–3236 (2020)
27. Xu, J., Tang, H., Ren, Y., Peng, L., Zhu, X., He, L.: Multi-level feature learning for contrastive multi-view clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16051–16060 (2022)
28. Yan, W., Zhang, Y., Lv, C., Tang, C., Yue, G., Liao, L., Lin, W.: Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19863–19872 (2023)
29. Yang, M., Li, Y., Hu, P., Bai, J., Lv, J., Peng, X.: Robust multi-view clustering with incomplete information. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 1055–1069 (2022)
30. Yang, M., Li, Y., Huang, Z., Liu, Z., Hu, P., Peng, X.: Partially view-aligned representation learning with noise-robust contrastive loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1134–1143 (2021)
31. Zeng, P., Li, Y., Hu, P., Peng, D., Lv, J., Peng, X.: Deep fair clustering via maximizing and minimizing mutual information: Theory, algorithm and metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23986–23995 (2023)
32. Zeng, P., Yang, M., Lu, Y., Zhang, C., Hu, P., Peng, X.: Semantic invariant multi-view clustering with fully incomplete information. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**, 2139–2150 (2023)
33. Zhou, R., Shen, Y.D.: End-to-end adversarial-attention network for multi-modal clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14619–14628 (2020)