Unified Representation Learning for Multi-View Clustering by Between/Within View Deep Majorization

Yue Zhang, Sirui Yang, Weitian Huang, Chang-Dong Wang, *Senior Member, IEEE*, and Hongmin Cai, *Senior Member, IEEE*

***Abstract*—Multi-view data are characterized by a variety of features that are semantically coherent, with different views usually complementing each other. Multi-view clustering seeks to learn a unified representation that is comprehensive yet discriminative to partition the samples. However, existing research ignored the possible redundancy of latent feature in each view, then yielded a unified task-oriented feature based on the correlation of these multi-view latent features without explicit consideration of the influence of the unified feature. Such a way failed to accurately measure the association among views. To fully take the association into account, we propose a deep multi-view Unified Representation Learning network, named by deep-URL, using measurement between multi-view data to realize desirable unified representation. Specifically, deep-URL employs a multi-view encoder-decoder framework that maps multi-scale data into a same latent space to obtain view-specific latent representation. Within each view, the view-specific features are pruned to eliminate the redundant information in the latent space by minimizing the within-view mutual information. Between the multiple views, a unified informative representation is derived to accurately estimate the nonlinear associations between views via maximizing conditional mutual information of the view-specific features. The deep unified representation is further elaborated to achieve the full reconstruction of each view to ensure its reliability. The proposed network is trained by between/within view deep representation majorization. It achieves redundant representation removal within each view, unified features integration across views, and universal reconstruction via decoding. Extensive experiments on five real datasets with largely varied sample sizes demonstrate that the deep-URL achieved superior clustering performance by comparing with fourteen baseline methods.**

***Index Terms*—Multi-view clustering, multi-view learning, unsupervised learning, conditional mutual information.**

I. Introduction

IN REAL-WORLD applications, observed samples have distinct properties that often reveal across and within modalities characterizes. For example, in clinical practice [1], written records and medical images on a patient reflect the patient's condition. In the computer vision community, various image descriptors such as SIFT [2], HOG [3] and GIST [4] are designed to measure different properties on interested object [5]. Massive previous works [6], [7], [8], [9], [10], [11] have empirically shown that multi-view data contains complementary information, leading to a comprehensive understanding for data.

To fully exploit the valuable structure information in multi-view data, many types of research have been proposed to develop multi-view clustering. Multi-view clustering research has introduced tools for data analysis from the perspective of Computational Intelligence. For example, following the trend of deep learning, multi-view clustering has increasingly involved neural networks for feature extraction and computation and expanded the application field of Computational Intelligence research.

A fundamental problem in multi-view clustering analysis is to explore potential complementary information and integrate it into a unified representation as the basis for grouping samples. A simple way is to directly concatenate multiple views ignoring the correlation between views, but due to the heterogeneous and high-dimensional nature of multi-view data, this may result in degraded clustering performance [12], [13]. In order to rationally integrate multi-view information, a significant amount of recent work has focused on learning a unified representation while taking into account inter-view correlations and information redundancy [14], [15], [16], [17]. A typical multi-view clustering method via canonical correlation analysis (CCA) [18] aims to learn the consistent representations by maximizing the linear correlation between views. On the other hand, in order to initially eliminate the large amount of redundant information brought by the high-dimensional raw data space, feature dimensionality reduction using deep neural networks (DNN) is widely used

Manuscript received 6 May 2023; accepted 22 June 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFE0112200, in part by the Key-Area Research and Development of Guangdong Province under Grants 2022A0505050014 and 2022B1111050002, in part by the Key-Area Research and Development Program of Guangzhou City under Grants 202206030009 and 2023B01J0002, in part by the National Natural Science Foundation of China under Grants U21A20520 and 62172112, and in part by Guangdong Key Laboratory of Human Digital Twin Technology under Grant 2022B1212010004. *(Corresponding author: Yue Zhang.)*

Yue Zhang is with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510640, China (e-mail: zhangyue@gpnu.edu.cn).

Sirui Yang, Weitian Huang, and Hongmin Cai are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: puremorning@yeah.net; cs_wthuang@mail.scut.edu.cn; hmcai@scut.edu.cn).

Chang-Dong Wang is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China (e-mail: changdongwang@hotmail.com).

Codes are available on https://github.com/PureRRR/deepURL.

Digital Object Identifier 10.1109/TETCI.2023.3314551

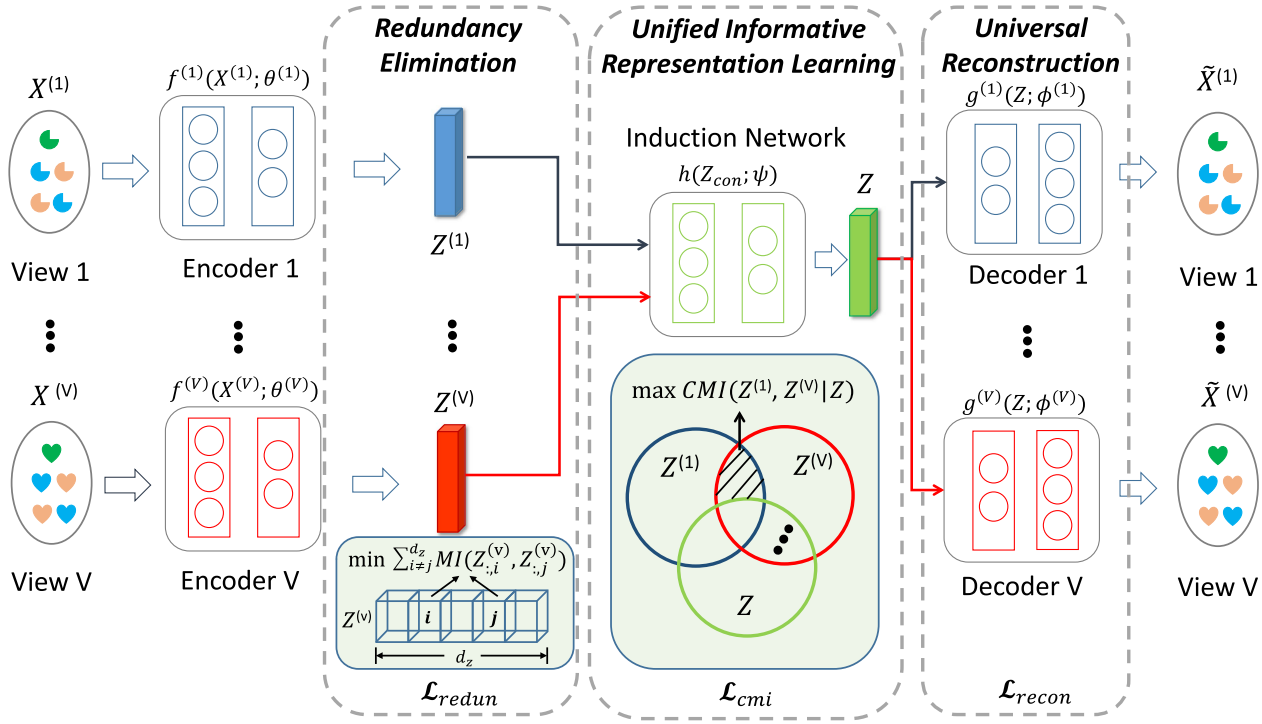


Fig. 1. Architecture of deep multi-view unified representation learning network (deep-URL). The framework cooperatively optimizes three components in a multi-view encoder-decoder structure, including redundancy elimination, unified informative representation learning and universal reconstruction.

by recent multi-view clustering methods. For example, deep canonical correlation analysis [19] and deep canonically correlated auto-encoders (DCCAE) [12], deep-based invariants of CCA, introduce the framework of encoders to map the raw multi-view data to a low-dimensional space and then perform CCA to obtain a unified latent representation. However, the dimensionality of the latent space is usually set empirically, which may result in low-dimensional features that are often highly correlated with each other. Such an approach is not an effective way to obtain concise latent representations. Worse still, CCA-based method is stuck in the dilemma of incapability to measure nonlinear correlation between views and suffers from the issue of less informative representation.

To address this shortcoming, the recently proposed multi-view contrastive learning approaches [20], [21], [22], [23], [24] utilize mutual information (MI) to estimate and maximize the nonlinear associations between the latent variables, followed by post-processing to obtain a unified representation. However, directly estimating the association between two latent variables without considering the existence of correlation with a third variable (the unified representation) may lead to an overestimation of mutual information [25], and results in a noncritical unified representation obtained in the end.

Besides, original data is often obtained without sufficient preprocessing, leading to redundant information between dimensions in an n-dimensional feature. This redundancy can negatively impact downstream representation learning. For example, Local Binary Pattern (LBP) feature is used for describing image texture and it is produced based on eight surrounding pixels, which inevitably retains some overlapping information. Existing research pay less attention to redundancy within latent representations, and such redundancy might be greatly magnified in the following unified representation, which results in unsatisfactory downstream performance.

In summary, learning a deep unified informative yet discriminative representation for multi-view clustering is still an unsolved problem. In this article, we present a deep multi-view Unified Representation Learning network from the view of information theory, namely deep-URL. It seeks to minimize redundant information in the latent space and maximize the correlation between views, then integrate discriminative information into a unified informative representation. The overall architecture of deep-URL is given in Fig. 1. To be specific, deep-URL first maps each view to a low-dimensional latent space on which the feature-wised mutual information is low-level to avoid overlap between features so that concise and sufficient information is contained in the limited dimensional space. The deep unified representation can then be obtained through the induction network, which is trained by maximizing the conditional mutual information of multiple latent variables from different views over the unified variable. The obtained unified features are then used by multiple decoders to reconstruct the multi-view sample, consistently ensuring that the identified unified features contain essential information for each view.

To summarize, the main contributions of our work are illustrated as follows,

- An information-aware deep multi-view learning approach is proposed to excavate a deep unified representation. The redundant information in the latent space is eliminated by minimizing the feature-wised mutual information of the

view-specific features, achieving the learning of concise representations.

- Deep-URL adopts conditional mutual information to avoid the overestimation of nonlinear associations between view-specific features, and to learn a deep unified representation in an informative yet discriminative latent space. To the best of our knowledge, we are the first to utilize conditional mutual information to estimate multi-view nonlinear correlations.
- Extensive experiments by comparing with fourteen popular methods demonstrate that the proposed method achieved superior clustering performance by a large margin.

II. Related Work

A. Multi-View Clustering

Multi-view clustering aims at revealing the inter-cluster and intra-cluster information and then partitioning samples by exploring underlying correlations and mining valuable content from different views. A better representation learned by comprehensive consideration of multiple views is crucial to multi-view clustering. Multi-view clustering can be realized by two mainstream approach, i.e., conventional methods and deep-based methods.

Co-training is one of the earliest schemes for learning in multi-view clustering. Given two distinctive views from unlabelled data, co-training works by alternately minimizing disagreement among multiple views. Several co-training-based variants have been proposed. Kumar et al. [26] revealed underlying consistent clusterings hidden in multi-view data through co-regularization process of clustering hypothses. Sindhwani et al. [27] proposed the construction of a data-dependent co-regularization norm, extending the existing algorithmic scope of co-regularization. CCA-based methods are another representative stream for multi-view clustering. Chaudhuri et al. [18] sought to find basis vectors for two sets of variables by alternately maximizing the correlations between the projections onto these basis vectors, which one can straightforwardly apply to bi-view data to reach a consensus.

On the other hand, due to the remarkable performance of deep models, multi-view paradigm has been increasingly adopted in many domains, including artificial intelligence and computer vision. For instance, the deep learning version of CCA, deep canonical correlation analysis [19] projects raw data into deep features by nonlinear transformation and its improved version deep canonically correlated auto-encoders (DCCAE) [12] adds auto-encoder to strengthen the representational ability. As a powerful unsupervised learning technique, generative adversarial networks [28] has been widely used and yielded promising results. Li et al. [29] propose to recover each view from the learned latent representation using adversarial learning. Zhang et al. [30] propose to generate the missing views according to available views and finally obtain a complete latent representation and nice performance in clustering. However, one of the drawbacks of generative adversarial networks (GANs) is that the training process might be unpredictably unstable and pose a challenge to the final convergence.

B. Multi-View Information Learning

Recently, some works propose to introduce information theory into multi-view learning. Typically, mutual information is leveraged for inter-view shared representation learning [21] recently and yields stable and nice performance. [31] proposed to involve mutual information constraint to minimize the common representations and view-specific representations for obtaining multi-level information, then a maximization process of mutual information is displayed on each instance and its k-nearest neighbors to boost intra-cluster aggregation. [20] proposed a self-supervised based representation learning method and utilized mutual information maximization between features from multiple views in a shared context for better learning ability. [32] proposed to apply information bottleneck principle and utilize mutual information to exclude irrelevant and noisy information for improved representation learning. [33] extended it to the unsupervised field and took advantage of self-supervision and data-augmentation. Although mutual information can measure nonlinear relationships and benefit common information discovery, it naturally lacks the ability to detect direct associations and results in wrongly estimated correlation, motivating the proposed deep-URL to adopt CMI-based learning for more accurate representation exploration.

III. Deep Multi-View Unified Representation Learning Network

To simplify the description, here we take the dual-view data as illustrative example. The case of more than two views is discussed and experimented (see Experiments section). Given two views of samples $\boldsymbol{X} = \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$, where $\boldsymbol{X}^{(v)} \in R^{n \times d_v}$ denotes v-th view data, with n and d_v representing the number of samples and the dimension of v-th view, respectively. We use deep encoder-decoder networks to extract the view-specific low-dimensional features $\{\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}\}$ for the two views. The two latent representations are learned such that the features should not be redundant with respect to each other, i.e., each individual feature within a view-specific latent representation should not be highly correlated. It is realized by minimizing view-specific feature-wised mutual information. Upon having the informative yet concise low-dimensional features from the two views, our ultimate objective is to learn a view-independent unified features $\boldsymbol{Z}$. The feature $\boldsymbol{Z}$ has two roles, it is highly unified in that it is correlated with both view-specific low-dimensional features $\{\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}\}$, and it is generative in that the unified features $\boldsymbol{Z}$ can be used to reconstruct the raw samples from each view. Accordingly, our objective is realized by a deep multi-view Unified Representation Learning network, illustrated in Fig. 1. The entire network is trained by collaboratively optimizing three components, including removing redundant view-specific features, deep unified representation learning, and reconstruction via the unified features.

A. Preliminaries

Given two random variables $\boldsymbol{X} \sim p(\boldsymbol{x})$ and $\boldsymbol{Y} \sim p(\boldsymbol{y})$, the amount of information they share can be measured in terms of

Shannon's Mutual Information (MI). For discrete distributions, the mutual information is calculated as,

$$MI(\boldsymbol{X};\boldsymbol{Y}) = \sum_{\boldsymbol{x},\boldsymbol{y}} p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) \log \frac{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{p_{\boldsymbol{X}}(\boldsymbol{x}) p_{\boldsymbol{Y}}(\boldsymbol{y})} \tag{1}$$

where $p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ is the joint probability distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$, $p_{\boldsymbol{X}}(\boldsymbol{x})$ and $p_{\boldsymbol{Y}}(\boldsymbol{y})$ are the marginal distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

MI can be used to measure nonlinear association between variables, in contrast to Pearson Correlation Coefficient, which measures linear association. MI is non-negative and its value is equal to zero if and only if the two variables are independent, while the larger the value of MI, the stronger the association of the two variables.

However, if two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both associated with a common random variable $\boldsymbol{Z}$, the mutual information $MI(\boldsymbol{X};\boldsymbol{Y})$ fails to characterize the direct association, which leads to overestimation. Accordingly, conditional mutual information is designed to fix the flaw, and it is defined by,

$$\begin{aligned} & CMI(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \\ &= \sum_{\boldsymbol{z}} p_{\boldsymbol{Z}}(\boldsymbol{z}) \sum_{\boldsymbol{x},\boldsymbol{y}} p_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z}) \log \frac{p_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})}{p_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y}|\boldsymbol{z})} \\ &= \sum_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} p_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \log \frac{p_{\boldsymbol{Z}}(\boldsymbol{z}) p_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z}) p_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y},\boldsymbol{z})} \end{aligned} \tag{2}$$

It is proved that when two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are dependent on a third variable $\boldsymbol{Z}$, for any joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$, there exists $MI(\boldsymbol{X};\boldsymbol{Y}) \geq CMI(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$ [25].

In this article, we introduce Fano's inequality to prove the rationality of measuring the redundancy within view-specific feature by mutual information.

Lemma 1: For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, Fano's inequality is concerned with the probability of error in the situation that we think of $\hat{X}$ as a "guess" for X, defined as $P_e = \mathbb{P}[X \neq \hat{X}]$. Then,

$$H(P_e) + P_e \log |\chi| \geq H(X|\hat{X}) \geq H(X|Y)$$

where $H(\cdot)$ denotes information entropy.

Proof: Define

$$E = \begin{cases} 0, & X = \hat{X} \\ 1, & X \neq \hat{X} \end{cases}$$

Using chain rule to expand $H(E, X|\hat{X})$ in two different ways,

$$\begin{aligned} H(E, X \,|\, \hat{X}) &= H(X|\hat{X}) + \underbrace{H(E|X,\hat{X})}_{=0} \\ &= \underbrace{H(E|\hat{X})}_{\leq H(E)=H(P_e)} + H(X|E,\hat{X}) \end{aligned}$$

Then,

$$\begin{aligned} H(X|E,\hat{X}) &= \mathbb{P}[E=0] H(X|\hat{X}, E=0) \\ &\quad + \mathbb{P}[E=1] H(X|\hat{X}, E=1) \\ &\leq (1-P_e)\cdot 0 + P_e \log_2 |\chi| \end{aligned}$$

Finally, we have

$$H(P_e) + P_e \log |\chi| \geq H(X|\hat{X}) \geq H(X|Y)$$

Since $H(P_e) \geq 0$, a corollary can be:

$$P_e \geq \frac{H(X|Y) - 1}{\log_2 |\chi|} = \frac{H(X) - I(X;Y) - 1}{\log_2 |\chi|}$$

The inequality states that for any estimator $\hat{X}$, the probability of prediction error from X to Y is lower bounded by an expression dependent on the mutual information $MI(X;Y)$. As the mutual information decreases, the bound is maximized; whether or not the bound can be reached depends on the ability of estimator $\hat{X}$.

B. Elimination of View-Specific Redundancy Features

We employ a deep multi-view encoder-decoder networks to obtain the low-dimensional features for each view. The low-dimensional features are non-unique in representing the view-specific characteristics. Besides, there might exist large redundant information that may hinder the learning of concise representation. In general, it is widely recognised that a good set of features should not only be individually relevant to the learning task, but also should not be redundant with respect to each other. The criteria based on information gain has long been used to extract a informative feature subset. Mathematically, for each view, the encoder learns a view-specific representation $\boldsymbol{Z}^{(v)} \in R^{n \times d_z}$ given $\boldsymbol{X}^{(v)}$ that downscales the original data dimension d_v to the dimension d_z of the same latent space. The feature extraction can be formulated by $\boldsymbol{Z}^{(v)} = f^{(v)}(\boldsymbol{X}^{(v)};\, \theta^{(v)})$, where $f^{(v)}$ represents v-th view's encoder parameterized by $\theta^{(v)}$. Here, a two-layer independent fully connected network and a softmax layer are implemented to achieve the goal. To eliminate the redundant features within the latent space, a mutual information loss is imposed for each view to reduce the redundant information within the latent representation $\boldsymbol{Z}^{(v)}$. It is not difficult to show that the Bayes error of predicting $\boldsymbol{Z}^{(v)}_{:,i}$ from $\boldsymbol{Z}^{(v)}_{:,j}$ is lower-bounded by Fano's inequality according to Lemma 1.

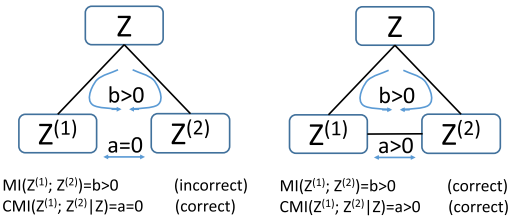
$$P\left(g\left(\boldsymbol{Z}^{(v)}_{:,i}\right) \neq \boldsymbol{Z}^{(v)}_{:,j}\right) \geq \frac{H\left(\boldsymbol{Z}^{(v)}_{:,i}\right) - MI\left(\boldsymbol{Z}^{(v)}_{:,i}; \boldsymbol{Z}^{(v)}_{:,j}\right) - 1}{\log\left(|\boldsymbol{Z}^{(v)}_{:,i}|\right)}$$

where $\boldsymbol{Z}^{(v)}_{:,i} \in R^{n\times 1}$ represents the i-th dimension of v-th view, $H(\boldsymbol{Z}^{(v)}_{:,i})$ is the information entropy of the random variable $\boldsymbol{Z}^{(v)}_{:,i}$.

Therefore, our task is to eliminate the feature redundancy such that their accumulated mutual information is minimized, as is formulated in (3),

$$\mathcal{L}_{redun} = \sum_{v=1}^{2} \sum_{i \neq j}^{d_z} MI\left(\boldsymbol{Z}^{(v)}_{:,i}; \boldsymbol{Z}^{(v)}_{:,j}\right) \tag{3}$$

Obviously, the feature pruning is independent of the network parameter $\theta^{(v)}$. Thus the obtained features are concise yet informative through reducing intra-view dependencies.



(a) Conditional independent (b) Direct dependent

Fig. 2. Differences between MI and CMI in both cases, when (a) $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ are conditionally independent given $\boldsymbol{Z}$, MI gives the wrong signal due to the influence of the common variable $\boldsymbol{Z}$. And when (b) $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ have direct dependency, both MI and CMI correctly measure the association between the two variables.

C. Learning a Unified Informative Features Via CMI

Our ultimate purpose is to learn a unified feature which contains coherent semantic context shared by multiple views, while being discriminative for clustering. The popular strategy to achieve the task is by multi-view contrastive learning [20], [21], [22] via maximizing the mutual information of two latent features. However, such a heuristic operation ignores the fact that the intrinsic correlation between view-specific latent variables is conditional on a shared variable. Accurate estimation of correlations between latent variables is impeded by $MI(\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)})$. As shown in Fig. 2(a), if two independent variables $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ are both related to a common variable $\boldsymbol{Z}$, the value of MI between $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ is larger than zero whereas the actual value should be zero since the two variables are conditionally independent. Only in the case of Fig. 2(b), the MI will correctly measure the association between $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ as in the case of CMI. This suggests that estimating the relationship between view-specific variables by MI can cause the problem of overestimation, i.e., $MI(\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}) \geq CMI(\boldsymbol{Z}^{(1)}; \boldsymbol{Z}^{(2)}|\boldsymbol{Z})$, which ultimately results in the acquisition of noncritical representations.

We now sought to maximize the conditional mutual information to seek unified features,

$$\mathcal{L}_{cmi} = -CMI\left(\boldsymbol{Z}^{(1)}; \boldsymbol{Z}^{(2)}|\boldsymbol{Z}\right) \tag{4}$$

where the unified features $\boldsymbol{Z}$ is obtained via a two-lay network that take view-specific latent representations as input,

$$\boldsymbol{Z}_{con} = concat\left(\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}\right),$$

$$\boldsymbol{Z} = Softmax\left(h\left(\boldsymbol{Z}_{con};\ \psi\right)\right)$$

where $concat(\cdot)$ denotes concatenation operation along feature dimension; $h(\cdot)$ denotes induction network with fully connected layers parameterized by ψ; $Softmax(\cdot)$ denotes a softmax layer; $\boldsymbol{Z}_{con} \in R^{n\times 2d_z}$, $\boldsymbol{Z} \in R^{n\times d_z}$.

Here, the features $\boldsymbol{Z}^{(1)}$, $\boldsymbol{Z}^{(2)}$ and $\boldsymbol{Z}$ are seen as a distribution of three discrete variables over d_z clusters. Their joint probability distribution $p_{\boldsymbol{Z}^{(1)},\boldsymbol{Z}^{(2)},\boldsymbol{Z}}$ is defined as $\mathcal{P} \in \mathcal{R}^{d_z\times d_z\times d_z}$,

$$\mathcal{P} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i^{(1)} \circ \boldsymbol{z}_i^{(2)} \circ \boldsymbol{z}_i \tag{5}$$

where $\circ$ denotes outer product, $\boldsymbol{z}_i^{(1)}, \boldsymbol{z}_i^{(2)}$ and $\boldsymbol{z}_i$ represent the i-th sample of $\boldsymbol{Z}^{(1)}$, $\boldsymbol{Z}^{(2)}$ and $\boldsymbol{Z}$, respectively.

Consequently, the conditional mutual information loss is calculated by,

$$\mathcal{L}_{cmi} = -\sum_{\boldsymbol{z}^{(1)},\boldsymbol{z}^{(2)},\boldsymbol{z}} \mathcal{P}_{\boldsymbol{Z}^{(1)},\boldsymbol{Z}^{(2)},\boldsymbol{Z}} \log \frac{\mathcal{P}_{\boldsymbol{Z}^{(1)},\boldsymbol{Z}^{(2)},\boldsymbol{Z}}\mathcal{P}_{\boldsymbol{Z}}}{\mathcal{P}_{\boldsymbol{Z}^{(1)},\boldsymbol{Z}}\mathcal{P}_{\boldsymbol{Z}^{(2)},\boldsymbol{Z}}} \tag{6}$$

where $\mathcal{P}_{\boldsymbol{Z}}$, $\mathcal{P}_{\boldsymbol{Z}^{(1)},\boldsymbol{Z}}$ and $\mathcal{P}_{\boldsymbol{Z}^{(2)},\boldsymbol{Z}}$ denote the marginal probability distribution.

D. Universal Reconstruction Via the Unified Informative Features

The unified features $\boldsymbol{Z}$ is a latent variable having the maximum conditional mutual information with the view-specific features $\boldsymbol{Z}^{(v)}$ for each view-v. Consequently, it can be used to reconstruct the observed samples $\boldsymbol{X}^{(v)}$. The reconstruction can be realized by multiple decoder networks,

$$\tilde{\boldsymbol{X}}^{(v)} = g^{(v)}\left(\boldsymbol{Z};\ \phi^{(v)}\right) \tag{7}$$

where $g^{(v)}$ denotes the decoder network parameterized by $\phi^{(v)}$ for v-th view.

Thus the reconstruction loss can be formulated by,

$$\mathcal{L}_{recon} = \sum_{v=1}^{2}\left\|\tilde{\boldsymbol{X}}^{(v)} - \boldsymbol{X}^{(v)}\right\|_F^2 \tag{8}$$

Finally, our network is guided by the three losses to complete the three tasks of removing redundant view-specific features, deep unified representation learning, and reconstruction via unified features simultaneously. The overall objective function is as follows,

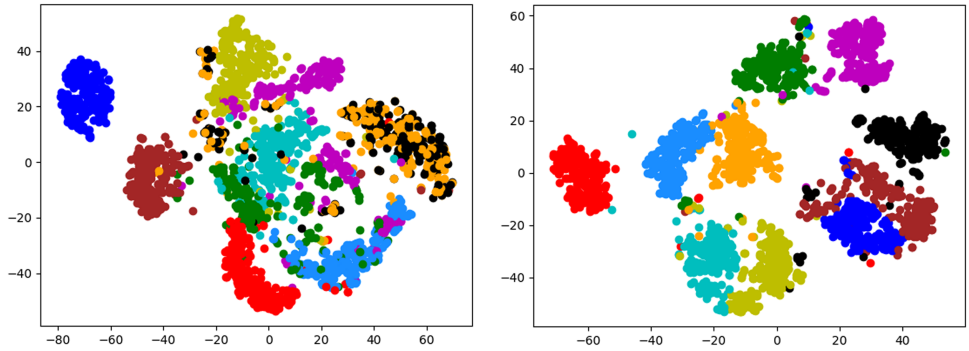
$$\begin{aligned}
&\min_{\theta^{(v)},\psi,\phi^{(v)}} \mathcal{L}_{redun} + \lambda_1\mathcal{L}_{cmi} + \lambda_2\mathcal{L}_{recon} \\
&= \sum_{v=1}^{2}\sum_{i\neq j}^{d_z} MI\left(\boldsymbol{Z}_{:,i}^{(v)}; \boldsymbol{Z}_{:,j}^{(v)}\right) \\
&\quad - \lambda_1 \sum_{\boldsymbol{z}^{(1)},\boldsymbol{z}^{(2)},\boldsymbol{z}} \mathcal{P}_{\boldsymbol{Z}^{(1)},\boldsymbol{Z}^{(2)},\boldsymbol{Z}} \log \frac{\mathcal{P}_{\boldsymbol{Z}^{(1)},\boldsymbol{Z}^{(2)},\boldsymbol{Z}}\mathcal{P}_{\boldsymbol{Z}}}{\mathcal{P}_{\boldsymbol{Z}^{(1)},\boldsymbol{Z}}\mathcal{P}_{\boldsymbol{Z}^{(2)},\boldsymbol{Z}}} \\
&\quad + \lambda_2 \sum_{v=1}^{2}\left\|\tilde{\boldsymbol{X}}^{(v)} - \boldsymbol{X}^{(v)}\right\|_F^2
\end{aligned} \tag{9}$$

where the hyperparameters λ_1 and λ_2 are the balanced constants.

The whole objective function is optimized using stochastic gradient descent and backpropagation algorithms.

E. Computational Analysis

The proposed method involves three parts for optimization and the total time complexity is equivalent to space complexity during running of programs. Given a dataset with sample size N, the maximal dimension across all views being d and the dimension of the learned representation being d_z, the complexity of calculating the redundancy elimination term is $O(d_z^2\ N)$. In



(a) View 1 (b) View 2

Fig. 3. Visualization by t-SNE on raw MNIST dataset. Clear chaos in the raw data can be observed.

TABLE I
DATASET SUMMARY

Dataset	Samples	View dimension	Classes
MSRC-V1	240	100, 256	7
MNIST	2000	76, 64	10
ANIMAL	10158	4096, 4096	50
ORL	400	4096, 4096	40
Caltech101-20	2386	512, 1984, 928	20

the second term of conditional mutual information, the complexity is $O(d_z^3 N + 2d_z^2\ N)$. Lastly, the complexity of universal reconstruction is $O(d^2\ N)$. To summarize, the total time and space complexity of the proposed model in each iteration is $O(d_z^3 N + d^2\ N)$.

IV. EXPERIMENTS

A. Evaluation Metrics

We adopt three classic metrics in clustering to measure the performance, including the normalized mutual information (NMI), accuracy (ACC) and Purity. For all of the metrics, a higher value indicates better performance. Note that the clustering results are obtained by conducting K-Means on the unified representation $\boldsymbol{Z}$.

B. Experimental Setup

Datasets: We extensively conduct experiments on five datasets to measure the effectiveness of the proposed method. Statistics on the tested datasets is summarized in Table I.

- *MSRC-V1* [34] is composed of 240 images with 9 object classes. We select 7 of the whole object classes, namely tree, building, airplane, cow, face, car and bicycle, and extract HOG, LBP features as 2 views.
- *MNIST* [35] is a dataset with 70000 handwritten digits with 28×28 pixels. In our experiment, we use a subset including 2000 samples with the flattened version of raw images and that of images with only edges existing acting as the input views.
- *ANIMAL* [36] contains 10158 images with animals divided into 50 categories. Two types of deep features are regarded as two views.
- *ORL*[1] consists of 400 images on 40 distinct subjects. Each subject is taken by ten different images with three descriptors including intensity, LBP and Gabor quantifying feature.
- *Caltech101-20* [37] is a subset of the object recognition dataset containing 20 classes with six different views, including histogram of oriented gradients, GIST features and local binary patterns.

Architectures and configurations: The encoders and decoders are constituted of the fully-connected architectures with softmax layer stacked over them. Specifically, the encoders have the same structure of *D-512-128-*d_z, where D denotes the dimension of inputs and d_z denotes the dimension of latent representation. The decoders are with a dimensionality of d_z*-128-512-1024-D*. For induction network, fully-connected network is also applied with the structure of $\sum d_z$*-128-128-*d_z, where $\sum d_z$ denotes the sum of the dimension of all views. Besides, the low-dimension d_z varies according to the applied dataset. In our experiments, for MSRC-V1, MNIST, ANIMAL, ORL and Caltech101-20, d_z is 64, 64, 128, 128 and 128, respectively.

Compared methods: A brief introduction of the compared methods is listed as follows.

A) Traditional methods

- *BestSV* [38] performs spectral clustering on each view of multi-view data and selects the best result among all views.
- *Co-reg* [26] co-regularizes clustering hypotheses to explore underlying consistent clusterings in multi-view data.
- *Co-pair* [26] encourages pairwise similarities under new representation based on the idea of Co-reg.
- *Diversity-induced Multi-view Subspace Clustering (DiMSC)* [39] introduces Hilbert Schmidt Independence Criterion (HSIC) for complementary information exploration.
- *Low-rank Tensor Constrained Multi-view Subspace Clustering (LT-MSC)* [40] seeks underlying high order correlations using low-rank constraint on tensors.
- *Consistent and Specific Multi-view Subspace Clustering (CSMSC)* [41] constructs self-representation property

[1]http://www.cl.cam.ac.uk/research/dtg/

TABLE II
CLUSTERING COMPARISONS ON FOUR BENCHMARKS

Methods		BestSv	Co-reg	Co-pair	DiMSC	CSMSC	RAMSC	DCCA	DCCAE	MvDSCN	DAMC	CDIMC	Completer	MvLNet	SURE	Ours
MNIST	ACC	0.582	0.710	0.720	0.320	0.849	0.782	0.726	0.756	0.817	0.755	0.857	0.732	0.871	0.892	**0.894**
	NMI	0.704	0.658	0.671	0.187	0.789	0.745	0.623	0.671	0.784	0.635	0.811	0.698	**0.857**	0.796	0.820
	PUR	0.640	0.731	0.727	0.326	0.849	0.782	0.610	0.631	0.817	0.755	0.868	0.744	0.867	0.778	**0.892**
MSRC-V1	ACC	0.600	0.572	0.705	0.708	0.671	0.690	0.592	0.611	0.605	0.510	0.738	0.535	0.633	0.681	**0.795**
	NMI	0.570	0.470	0.603	0.582	0.581	0.639	0.497	0.490	0.529	0.435	0.662	0.483	0.609	0.605	**0.745**
	PUR	0.600	0.598	0.713	0.709	0.710	0.719	0.483	0.477	0.605	0.543	0.736	0.535	0.620	0.505	**0.790**
ANIMAL	ACC	0.127	0.454	0.575	0.622	0.632	0.640	0.432	0.435	0.593	0.284	0.411	0.513	0.584	0.370	**0.646**
	NMI	0.207	0.567	0.694	0.712	0.720	0.749	0.311	0.313	0.703	0.485	0.701	0.601	0.700	0.504	**0.785**
	PUR	0.149	0.500	0.635	0.669	0.688	0.700	0.328	0.331	0.670	0.354	0.422	0.540	0.580	0.282	**0.720**
ORL	ACC	0.348	0.668	0.672	0.757	0.440	0.743	0.624	0.669	0.593	0.377	0.578	0.419	0.700	0.590	**0.775**
	NMI	0.592	0.833	0.834	0.880	0.653	0.877	0.510	0.573	0.703	0.521	0.730	0.702	0.688	0.450	**0.884**
	PUR	0.356	0.708	0.712	0.782	0.688	0.780	0.473	0.533	0.670	0.396	0.602	0.434	0.694	0.580	**0.797**

The optimal and suboptimal results are in bold and underlined, respectively.

through jointly utilizing common and unique information in each view.

- *Robust Auto-weighted Multi-view Subspace Clustering (RAMSC)* [42] learns appropriate weights for all views to obtain a similarity matrix, with a sparsity norm making algorithm robust.

B) Deep methods

- *Deep Canonical Correlation Analysis (DCCA)* [19] extract nonlinear deep features and apply CCA on these features to obtain common representations.
- *Deep Canonically Correlated Auto-Encoders (DCCAE)* [12] introduces an auto-encoder architecture to acquire deep features based on basic CCA.
- *Multi-view Deep Subspace Clustering Network (MvDSCN)* [43] leverages several deep diverse and universe networks to respectively extract individual and shared information.
- *Deep Adversarial Multi-view Clustering Network (DAMC)* [29] also applies auto-encoders to get deep features and additionally utilizes adversarial learning to keep the reconstructed data the same as input data.
- *Cognitive Deep Incomplete Multi-view Clustering Network (CDIMC)* [44] leverages graph embedding strategy and self-paced learning to yield high-level features and avoid influence of outliers.
- *Completer* [21] introduces maximization of mutual information between views and minimization of conditional entropy of different views to achieve dual prediction and representation learning.
- *Multi-view Laplacian Network (MvLNet)* [45] is the first deep version of the multi-view spectral representation learning method. MvLNet solves the problem of trivial solution when Laplacian embedding is simply combined with neural networks by posing an orthogonal restriction and transforming it into a deep layer.
- *Robust Multi-view Clustering with Incomplete Information (SURE)* [46] solves partially View-unaligned problem (PVP) and partially sample-missing problem (PSP) in a unified network by adopting contrasting learning paradigm and reduces the influence caused by random sampling through noisy-robust contrastive loss.

C. Experimental Results

1) Comparing With Baselines: In this section, we conduct clustering experiments to test the proposed network on four real databases by comparing with seven traditional methods and six deep methods. The experimental results are shown in Table II. It can be clearly seen that clustering results performed by multi-view methods in both traditional and deep stream uniformly outperform that by BestSV, i.e., single-view clustering, undoubtedly proving the necessity and effectiveness of multi-view clustering for revealing the underlying cluster information. Compared with other six traditional clustering methods, our method surpasses them by a large margin. For instance, on MSRC-V1, our method has a growth by 5.7%, 8.3% and 5.4% against the method reaches the second highest score with respect to ACC, NMI and Purity. The key reason lies in that traditional methods are greatly limited by using shallow and linear embedding functions, which are not able to capture the complex properties of real-world multi-view data. On the other hand, our method shows evident superiority compared with other deep methods. For instance, on MNIST, our method has a growth by 3.7%, 0.9% and 2.4% against the method reaching the second highest score with respect to ACC, NMI and Purity. Particularly, our method outperforms mutual-information-based method Completer with a clear improvement on four datasets. We attribute this improvement to the involvement of conditional mutual information for accurate correlation estimation.

2) Visualization: We intuitively analyzed the experimental results by visualizing the low-dimensional comprehensive and corresponding latent representation learned by our method and compared the method. The results on MNIST dataset is visualized via t-SNE in Fig. 4. Note that each kind of color denotes one category. The 2D visualization on the learned representations after DCCAE, MvDSCN, DAMC, CDIMC, Completer, MvLNet, SURE and our method are shown in Fig. 4(a)–(i), respectively. It can be clearly found that latent representations learned by DCCAE and MvDSCN can not be easily distinguished, especially between orange and black points. Although data points in DAMC and CDIMC mostly surround a specific center according to their categories, fuzzy boundaries indicate representations not discriminative enough. Likewise, Completer, MvLNet and SURE yield relatively clear boundaries but distribute the representations in a scattered manner. It is clearly shown that our method pushes homogeneous samples to assemble in a denser way and inhomogeneous samples to be separated in a farther way than by others. Both of the two observations prove that three key modules, i.e., within-view feature extraction with redundancy eliminated, deep unified representation learning and universal reconstruction, guide the model towards a clustering-friendly orientation and finally help to build a

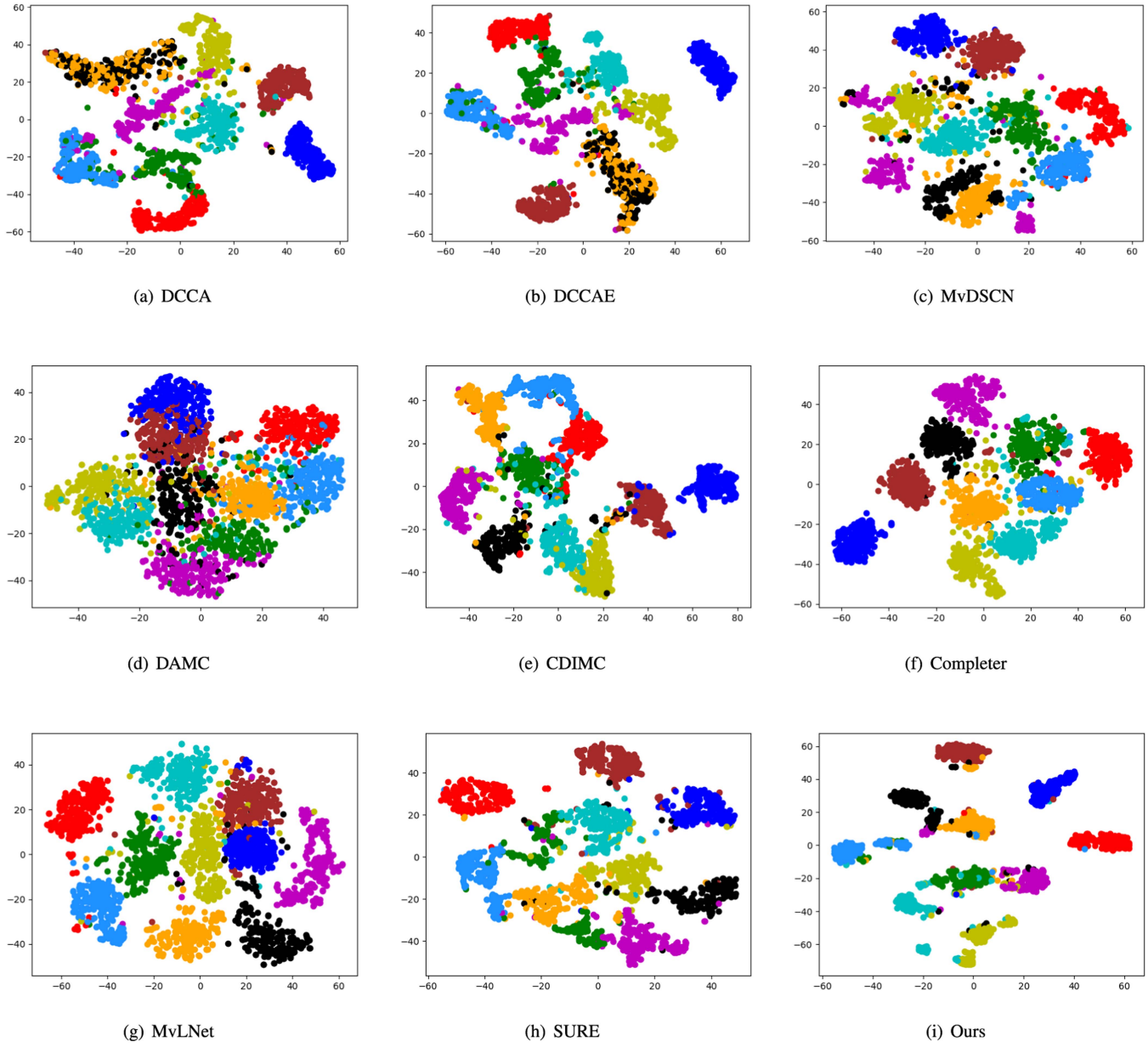


Fig. 4. Visualization to show the learned representations by t-SNE on MNIST dataset. More compact clusters and a greater distance between clusters are displayed by ours than by the others.

model that produces the unified informative and discriminative representation.

3) Ablation Study: To demonstrate the capability of redundancy elimination, CMI-based intrinsic representation learning module and universal reconstruction, we perform extensive ablation studies on four datasets. Table III reports the results on four datasets. $\mathrm{M}_{\sim redun}$ denotes the model without redundancy elimination on the basis of our model; $\mathrm{M}_{\sim recon}$ replaces the universal reconstruction with perspective reconstruction, i.e., the recovered sample is obtained by $\tilde{\boldsymbol{X}}^{(v)} = g^{(v)}(\boldsymbol{Z}^{(v)}; \phi^{(v)})$. $\mathrm{M}_{\sim cmi}$ removes the guidance of CMI. Instead, a MI maximization process is adopted and the comprehensive representation $\boldsymbol{Z}$ is acquired by averaging view-dependent representation $\boldsymbol{Z}^{(v)}$.

TABLE III
ABLATION STUDY ON FOUR BENCHMARKS

Methods		$\mathrm{M}_{\sim redun}$	$\mathrm{M}_{\sim recon}$	$\mathrm{M}_{\sim cmi}$	Ours
MNIST	ACC	0.879	0.874	0.870	**0.894**
	NMI	0.808	0.804	0.798	**0.820**
	PUR	0.870	0.874	0.858	**0.892**
MSRC-V1	ACC	0.752	0.742	0.741	**0.795**
	NMI	0.740	0.728	0.726	**0.745**
	PUR	0.751	0.750	0.740	**0.790**
ANIMAL	ACC	0.632	0.634	0.622	**0.646**
	NMI	0.770	0.772	0.755	**0.785**
	PUR	0.662	0.666	0.686	**0.720**
ORL	ACC	0.749	0.741	0.744	**0.775**
	NMI	0.870	0.870	0.862	**0.884**
	PUR	0.755	0.751	0.740	**0.797**

The optimal and suboptimal results are in bold and underlined, respectively.

TABLE IV
TRI-VIEW CLUSTERING COMPARISONS ON CALTECH101-20

Methods	BestSv	Co-reg	Co-pair	DiMSC	CSMSC	RAMSC	DCCA	DCCAE	MvDSCN	DAMC	CDIMC	Completer	MvLNet	SURE	Ours
ACC	0.604	0.404	0.420	0.280	0.414	0.488	0.382	0.364	0.538	0.357	0.556	0.452	0.520	<u>0.557</u>	**0.610**
NMI	0.537	0.577	0.568	0.342	0.630	<u>0.659</u>	0.339	0.339	0.642	0.490	0.634	0.424	0.620	**0.670**	**0.670**
PUR	0.664	0.747	0.744	0.571	0.694	<u>0.767</u>	0.673	0.669	0.548	0.670	0.572	0.460	0.518	0.562	**0.777**

The optimal and suboptimal results are in bold and underlined, respectively.

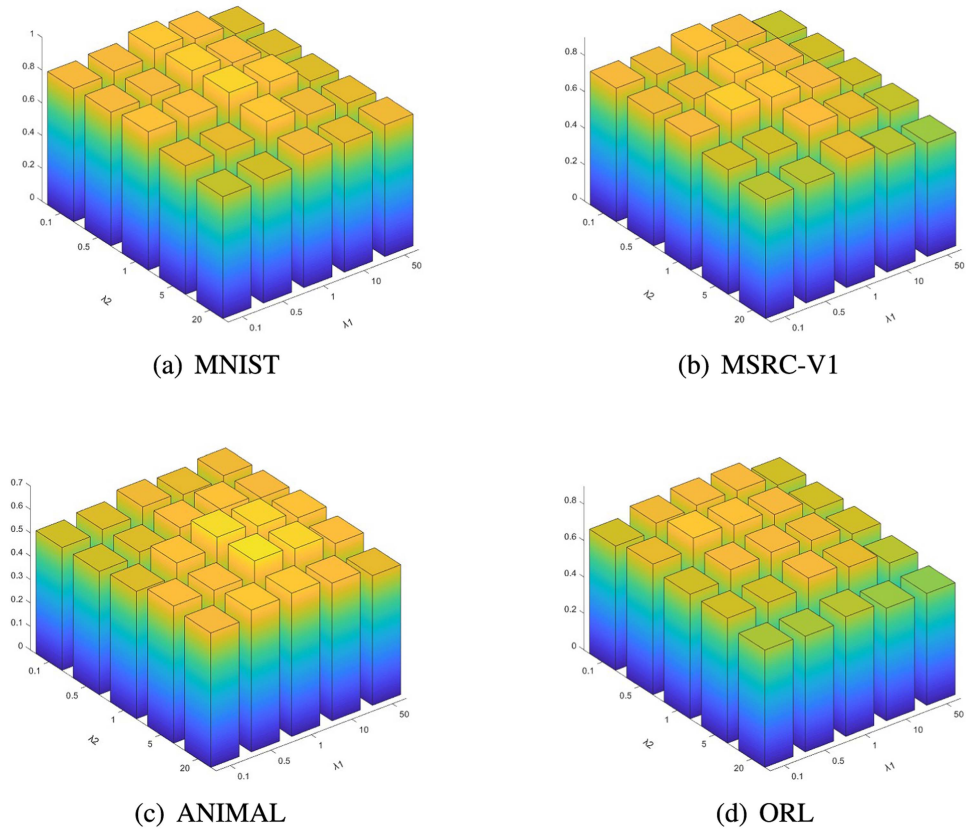


Fig. 5. Parameter analysis for ACC metric on four datasets. Evident peaks can be oberseved for four datasets and slight changes of scores are presented within given intervals.

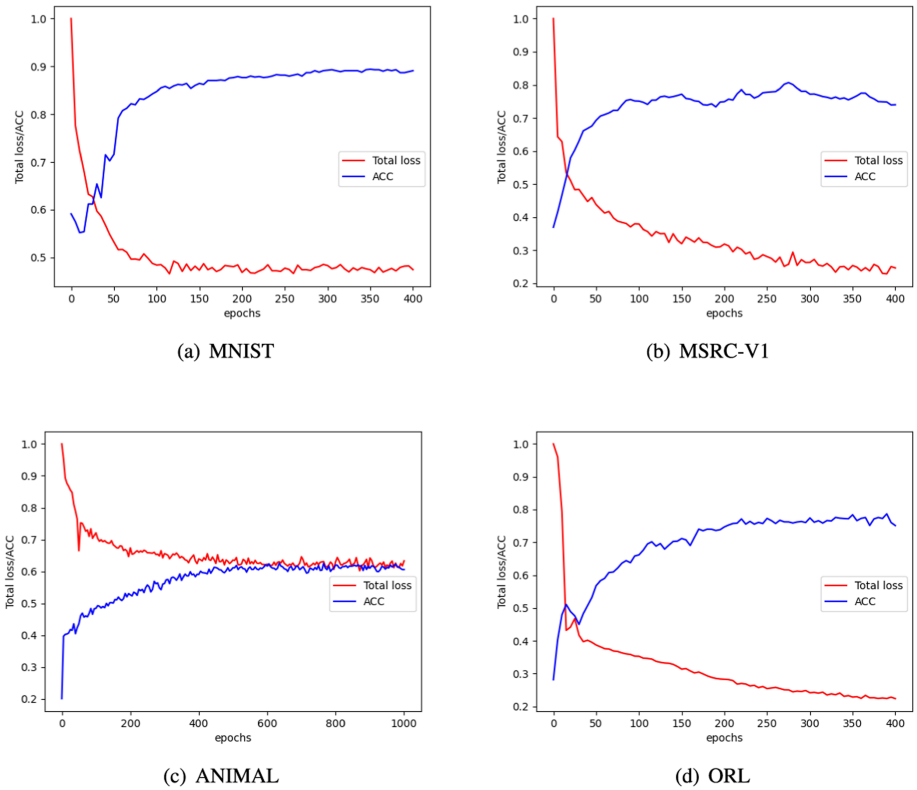
As can be observed from the results in Table III, removing any one of the components of the model results in a decrease in different degrees. Particularly, $\mathrm{M}_{\sim cmi}$ suffers much more decrease than $\mathrm{M}_{\sim redun}$ and $\mathrm{M}_{\sim recon}$. These results convey that every single module, especially CMI-guided learning, plays a crucial part when mining underlying relationships among multiple views.

4) Application to Tri-View Data: In this part, we conduct experiments on a tri-view dataset, Caltech101-20, to validate the availability of deep-URL. According to the definition of CMI presented in Equation 2, bi-view data correlation maximization can be realized through maximizing Equation 2. However, Equation 2 can not handle it when data contains more than two views unless one introduces interaction information, i.e., the generalization of mutual information. But another problem is that the definition of interaction information for three views involves a joint probability distribution of four variables, which produces a four-order tensor and requires extremely large memory. Thus, we turn to a practicable way, that is, replacing the CMI loss with the following form,

$$\begin{aligned}\mathcal{L}_{cmi} = & -CMI\left(\boldsymbol{Z}^{(1)};\boldsymbol{Z}^{(2)}|\boldsymbol{Z}\right) - CMI\left(\boldsymbol{Z}^{(1)};\boldsymbol{Z}^{(3)}|\boldsymbol{Z}\right) \\ & - CMI\left(\boldsymbol{Z}^{(2)};\boldsymbol{Z}^{(3)}|\boldsymbol{Z}\right) \end{aligned} \tag{10}$$

As is shown in Table IV , the performance of our method outperforms any other methods in traditional and deep stream in all metrics, indicating the applicability in multi-view data of our method. Note that Completer is removed from comparing method for the sake of unavailability of tri-view data.

5) Hyperparameter Analysis: To illustrate the impact of hyperparameters and the generalization of the proposed method, we present ACC with the hyperparameters varying on four benchmarks, as shown in Fig. 5. Each figure is plotted when two hyperparameters vary in specific intervals. We can clearly find that the proposed method deep-URL is insensitive to the tuning of hyperparameters on all datasets.



(a) MNIST (b) MSRC-V1

(c) ANIMAL (d) ORL

Fig. 6. Convergence analysis on four datasets. A clear tendency of convergence can be observed on four datasets for loss and ACC curve.

6) Convergence Analysis: To empirically analyze the convergence performance of the proposed deep-URL, we report the dynamic progress of total loss and ACC score during training on MNIST, MSRC-V1, ANIMAL and ORL. The total loss and ACC score are both normalized into the interval of [0, 1]. As is depicted in Fig. 6, both overall loss and ACC score fluctuate in the early stage but keep an evident trend of decreasing or increasing. Additionally, after around 400, 400, 1000 and 400 epochs, both total loss and ACC score reach a clearly stable state for MNIST, MSRC-V1, ANIMAL and ORL respectively, which shows nice convergence of the loss function.

V. Conclusion

This article proposes a deep multi-view unified representation learning network (deep-URL) to achieve effective multi-view clustering. It seeks to minimize the redundancy of the latent space and maximize the inter-view correlation, and then integrate view-specific features into a unified informative representation. The proposed method maps each view to a low-dimensional latent space, within which the concise yet informative view-specific features are obtained via minimizing feature-wised mutual information. To accurately estimate the nonlinear associations between views, we explicitly construct view-specific variables conditional on a unified variable and then maximize their conditional mutual information to learn a unified informative representation. By excavating the view-specific features within each view and coherent semantic features between views jointly, a unified set of features is obtained that can be used to completely reconstruct each view. The network is trained by jointly achieving redundancy removal within each view, unified features integration between views, and universal reconstruction via decoding until convergence. Experiments on five benchmarks empirically validate the superiority of our method compared with the state-of-the-art approaches.

References

[1] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn, "DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2103–2112.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[4] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[5] H. Zhao, Z. Ding, and Y. Fu, "Multi-view clustering via deep matrix factorization," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2921–2927.

[6] L. Fu, P. Lin, A. V. Vasilakos, and S. Wang, "An overview of recent multi-view clustering," *Neurocomputing*, vol. 402, pp. 148–161, 2020.

[7] D. J. Trosten, S. Lokse, R. Jenssen, and M. Kampffmeyer, "Reconsidering representation alignment for multi-view clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1255–1265.

[8] W. Xia, Q. Wang, Q. Gao, X. Zhang, and X. Gao, "Self-supervised graph convolutional network for multi-view clustering," *IEEE Trans. Multimedia*, vol. 24, pp. 3182–3192, 2022.

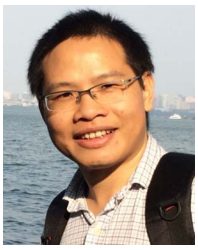
[9] X. Fang, Y. Hu, P. Zhou, and D. O. Wu, "Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 4, pp. 913–927, Aug. 2022.

[10] Y. Wu et al., "Multi-view point cloud registration based on evolutionary multitasking with bi-channel knowledge sharing mechanism," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 2, pp. 357–374, Apr. 2023.
[11] J. Yang and C.-T. Lin, "Multi-view adjacency-constrained hierarchical clustering," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 4, pp. 1126–1138, Aug. 2022.
[12] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.
[13] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 1863–1883, Oct. 2019.
[14] X. Li, H. Zhang, R. Wang, and F. Nie, "Multiview clustering: A scalable and parameter-free bipartite graph fusion method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 330–344, Jan. 2022.
[15] C. Zhang et al., "Generalized latent multi-view subspace clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 86–99, Jan. 2020.
[16] M. Soltanolkotabi et al., "Robust subspace clustering," *Ann. Statist.*, vol. 42, no. 2, pp. 669–699, 2014.
[17] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.
[18] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. 26th Int. Conf. Mach. Learn.*, 2009, pp. 129–136.
[19] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[20] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 15535–15545.
[21] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, and X. Peng, "Completer: Incomplete multi-view clustering via contrastive prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11174–11183.
[22] K. Do, T. Tran, and S. Venkatesh, "Clustering by maximizing mutual information across views," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9928–9938.
[23] M. Wu, S. Pan, and X. Zhu, "Attraction and repulsion: Unsupervised domain adaptive graph contrastive learning network," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 5, pp. 1079–1091, Oct. 2022.
[24] Z. Yuan, G. Li, Z. Wang, J. Sun, and R. Cheng, "RL-CSL: A combinatorial optimization method using reinforcement learning and contrastive self-supervised learning," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 4, pp. 1010–1024, Aug. 2023.
[25] J. Zhao, Y. Zhou, X. Zhang, and L. Chen, "Part mutual information for quantifying direct associations in networks," *Proc. Nat. Acad. Sci.*, vol. 113, no. 18, pp. 5130–5135, 2016.
[26] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1413–1421.
[27] V. Sindhwani, P. Niyogi, and M. Belkin, "A co-regularization approach to semi-supervised learning with multiple views," in *Proc. ICML Workshop Learn. With Mult. Views*, 2005, pp. 74–79.
[28] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[29] Z. Li, Q. Wang, Z. Tao, Q. Gao, and Z. Yang, "Deep adversarial multi-view clustering network.," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 2952–2958.
[30] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu, "Deep partial multi-view learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2402–2415, May 2022.
[31] F. Lele, Z. Lei, W. Tong, C. Chuan, Z. Chuanfu, and Z. Zibin, "Multi-view clustering from the perspective of mutual information," 2023, *arXiv:2302.087432*023.
[32] Q. Wang, C. Boudreau, Q. Luo, P.-N. Tan, and J. Zhou, "Deep multi-view information bottleneck," in *Proc. SIAM Int. Conf. Data Mining*, 2019, pp. 37–45.
[33] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, "Learning robust representations via multi-view information bottleneck," in *Proc. Int. Conf. Learn. Representations*, 2020.
[34] J. Winn and N. Jojic, "Locus: Learning object classes with unsupervised segmentation," in *Proc. IEEE/CVF 10th Int. Conf. Comput. Vis.*, 2005, pp. 756–763.
[35] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 469–477.
[36] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 951–958.
[37] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshop*, 2004, pp. 178–178.
[38] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 849–856.
[39] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang, "Diversity-induced multi-view subspace clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 586–594.
[40] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao, "Low-rank tensor constrained multiview subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1582–1590.
[41] S. Luo, C. Zhang, W. Zhang, and X. Cao, "Consistent and specific multi-view subspace clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3730–3737.
[42] W. Zhuge, C. Hou, Y. Jiao, J. Yue, H. Tao, and D. Yi, "Robust auto-weighted multi-view subspace clustering with common subspace representation matrix," *PLoS One*, vol. 12, no. 5, pp. 1–20, 2017.
[43] P. Zhu, B. Hui, C. Zhang, D. Du, L. Wen, and Q. Hu, "Multi-view deep subspace clustering networks," 2019, *arXiv:1908.01978*.
[44] J. Wen, Z. Zhang, Y. Xu, B. Zhang, L. Fei, and G.-S. Xie, "CDIMC-Net: Cognitive deep incomplete multi-view clustering network," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2021, pp. 3230–3236.
[45] Z. Huang, J. T. Zhou, H. Zhu, C. Zhang, J. Lv, and X. Peng, "Deep spectral representation learning from multi-view data," *IEEE Trans. Image Process.*, vol. 30, pp. 5352–5362, 2021.
[46] M. Yang, Y. Li, P. Hu, J. Bai, J. C. Lv, and X. Peng, "Robust multi-view clustering with incomplete information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1055–1069, Jan. 2023.

Yue Zhang received the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, China, in 2017. She is currently an Associate Professor with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China. Her research interests include bioinformatics and Big Data mining.

Sirui Yang received the bachelor's degree in computer science from the China University of Mining Technology, Xuzhou, China. He is currently working toward the master's degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His research interests include data mining and multi-view clustering.

Weitian Huang received the M.S. degree in control science and engineering from the School of Automation, Guangdong University of Technology, Guangzhou, China, in 2020. He is currently working toward the Ph.D. degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou. His research interests include multi-view learning, deep generative model, clustering, and bioinformatics.



Chang-Dong Wang (Senior Member, IEEE) received the Ph.D. degree in computer science from Sun Yat-sen University, Guangzhou, China, in 2013. From January 2012 to November 2012, he was a Visiting Student with the University of Illinois at Chicago, Chicago, IL, USA. In 2013, he was an Assistant Professor with the School of Mobile Information Engineering, Sun Yat-sen University, where he is currently an Associate Professor with the School of Data and Computer Science. His research interests include machine learning and data mining.

Hongmin Cai (Senior Member, IEEE) received the B.S. and M.S. degrees in mathematics from the Harbin Institute of Technology, Harbin, China, in 2001 and 2003, respectively, and the Ph.D. degree in applied mathematics from The Hong Kong University, Hong Kong, in 2007. He is currently a Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. From 2005 to 2006, he was a Research Assistant with the Center of Bioinformatics, Harvard University, and Section for Biomedical Image Analysis, University of Pennsylvania. His research interests include biomedical image processing and omics data integration.