Deep Mutual Information Decoupling Based Unsupervised Image Clustering

Yanfeng Wang[1], Jinfeng Wang[2], and Weirong Zhang[2*]

[1] School of Mechanical and Electrical Engineering, Weifang Vocational College, Weifang, 262737, China
[2] School of Information Engineering, Weifang Vocational College, Weifang, 262737, China
[*] Corresponding author. E-mail: weirongzhang_wvc@163.com

Received: Jul. 14, 2024; Accepted: Aug. 24, 2024

Cross-view image clustering (CIC) showcases immense potential in recognizing image patterns due to the power to aggregate information between views without labels. However, most CIC ignore intricate coupling relationships between category and redundant features in aggregating complementary information of cross-view images, which may restrict performance in recognizing patterns of images. To this end, a cross-view mutual information decoupling based deep generative clustering approach is proposed for recognizing image patterns (DMID-UIC), which contains information maximization deep generative module and self-supervised posterior inference module. Specifically, the former maximizes the mutual information between data and semantics within the generative adversarial network to decouple category and redundant features hidden in cross-view images, which guarantees the separation of different semantic distributions in the data space. The latter models the distribution fitting between generated data and the prior semantic code as a classification task, via treating partitioning results of common representations between views as self-supervised labels. Meanwhile, to better optimize the model, an EM optimization strategy is designed to enhance the above two module learning in an iterative manner. Finally, comprehensive results verify the superiority and effectiveness of DMID-UIC. DMID-UIC improves ACC by 10.1% on the Caltech 101-7 dataset, compared to the second-best result.

Keywords: Unsupervised image clustering; mutual information decoupling; deep cross-view learning

http://dx.doi.org/10.6180/jase.202506_28(6).0014

1. Introduction

Image clustering has been a fundamental and critical task in the field of computer vision since its inception, aiming to automatically group large amounts of image data based on visual content or semantic relevance [1, 2]. Early clustering methods relied primarily on manually extracted features, such as color, texture, and shape, which, while providing some visual differentiation, often failed to deeply reveal the deeper semantics of images. As technology evolved, clustering algorithms began to attempt the combination of various features to enhance the accuracy and robustness of clustering. However, real-world image data is far more complex than initially recognized. Images are not just static collections of pixels; they are often associated with rich contextual information, such as the shooting environment, related text descriptions, audio information, etc. These sources of information, which we refer to as views provide new perspectives and possibilities for image clustering [3, 4]. Against this backdrop, cross-view clustering has emerged, integrating information from different views to provide a more comprehensive feature representation for image clustering. The core advantage of cross-view clustering lies in its ability to leverage the complementarity between different views to improve the accuracy and robustness of clustering. For example, visual information may be affected by lighting or occlusion in some cases, while text descriptions can provide additional semantic

information to assist in clustering. The development of cross-view clustering marks a shift in image clustering technology from single-dimensional feature analysis to a multi-dimensional, multi-perspective comprehensive analysis, providing more advanced and effective means for the organization and management of image data, and demonstrating broad application potential in fields [5, 6].

Cross-view image clustering (CIC), as a brilliant unsupervised cross-view image learning paradigm, aims to aggregate information scattered in heterogeneous views to mine intrinsic patterns of image. Recently, some deep cross-view image clustering methods combine deep neural networks with cross-view image clustering to explore image patterns from nonlinear fusion information, which achieves the state-of-the-art performance. They be roughly separated into two classes, i.e., deep fusion methods [7, 8] and deep clustering methods [4, 6]. The former utilizes self-supervision information of data to extract generalized fusion representations to characterize fusion information, and then performs vanilla single-view clustering methods to mine image patterns. This two-stage learning strategy may disconnect the fusion representation learning and clustering partition, causing that the fusion representations are not optimal for the clustering partition. To close the gap between the two processes, the deep clustering methods combine fusion representation learning and clustering partition of images into the joint optimization strategy. They utilize the divergences of data structures to guide the fusion of complementary information among views, as well as mining clustering patterns of images in an end-to-end manner.

Despite prior deep cross-view methods have demonstrated promising performance in recognizing image patterns, there still exists a significant challenge. Most deep cross-view methods do not take into full consideration correlations between data and clusters in aggregating consistent and complementary information of cross-view images, which leads to the suboptimal clustering performance in recognizing image patterns. In other words, they ignore intricate coupling relationships between cluster-related and redundant features in data and struggle to mitigate the interference of redundant information on the discriminative ability of common representations, degrading semantic robustness of common representations. Thus, they may fail in recognizing image patterns.

To address this challenge, a novel deep mutual information decoupling based generative cross-view representation learning method is proposed for unsupervised image clustering (DMID-UIC), consisting of an information maximization deep generative module and a self-supervised posterior inference module. The former models the joint probability distribution between data and clusters to generate cross-view images via view-specific generative adversarial networks with cluster priors, and then maximizes the mutual information between generation images and clusters to decouple clustering features and redundant features in cross-view data, which guarantees the separation of different cluster distributions in the data space. The latter defines the distribution fitting between generated data and the prior cluster information in the latent space as a classification task, via treating clustering results of common representations between views predicted by the clustering network as self-supervised labels. Meanwhile, to obtain the optimal parameters of the model and the optimal clustering posterior distribution, an EM optimization strategy is utilized to iteratively promote the information maximization deep generative learning and the self-supervised posterior inference learning. Finally, comprehensive qualitative and quantitative experiment results verify the superiority and effectiveness of DMID-UC.

The main contributions of DMID-UIC are threefold:

- A mutual information decoupling based generative cross-view representation learning is proposed via maximizing the mutual information between cross-view images and clusters in the generative process to achieve decoupling between clustering features and redundant features, which captures a clustering structure with clear boundaries for recognizing image patterns.

- An EM optimization strategy is designed to enhance the information maximization deep generative learning and the self-supervised posterior inference learning in an iterative manner, which effectively models the joint distribution of cross-view data and clusters.

- Comprehensive qualitative and quantitative experiments are designed on five real world datasets, the results verify that the superiority of DMID-UIC compared with eight cross-view methods, which establishes a fresh state-of-the-art benchmark for recognizing image patterns.

The subsequent sections of DMID-UIC are conduced as follows: section 2 shows the description of cross-view image clustering techniques. section 3 elaborates an in-depth exposition of the DMID-UIC. section 4 offers a comprehensive depiction of experiment evaluations. Lastly, section 5 concludes DMID-UIC.

2. Related works

DMID-UIC learns intrinsic patterns of cross-view images, which is closely related to deep fusion methods and deep clustering methods.

Deep fusion methods, as the two-stage methods, focus on extracting generalized fusion representations of cross-view images, and then leveraging a single-view clustering algorithm to mine clustering patterns. For example, Zhang et al. [9] learned collaboratively consistent and complementary information among views in common and specific latent spaces by fitting the degradation process from common representations to cross-view images using a nested encoding-decoding framework. They then performed k-means to explore data structures [9]. Zhang et al. [9] learned common representations with robustness by maximizing mutual information between view-common and view-specific representations and minimizing mutual information between cross-view data and view-specific representations. They then conducted the k -means algorithm to measure similarities between data for mining patterns [10]. Gao et al. [11] fused complementary information based on a consensus subspace that maximized correlations of views and then performed spectral clustering on the self-expression consensus matrix to mine clustering patterns [11]. Yang et al. [12] proposed a robust cross-view autoencoder architecture by designing a contrastive relaxation loss that alleviated the mutual exclusion of data within the same cluster, enhancing the aggregation of complementary information in heterogeneous views. This approach utilized non-linear semantic neighbors of data to impute view-missing representations for pattern mining [12]. DMCAG defined an anchor cross-view subspace clustering method to learn the consensus matrix between views and then employed spectral self-supervised learning to align view-specific clustering assignments for data pattern mining [13]. VCGA devised an information decoupled cross-view graph clustering method by exploring view-specific graphs and consensus graphs with local and global structure exploration, fully examining the local manifold to mine patterns of images [14].

Deep clustering methods integrate complementary information fusion with clustering pattern recognition in an end-to-end architecture, where divergences of data structures are used to super vise model training. Trosten et al. [15] utilized cross-view contrastive learning followed by adaptive linear combination to fuse complementary and consistent information between views. They then conducted Cauchy-Schwarz divergence in the fusion representation space to capture structure divergences between data for mining clustering patterns [15]. Xu et al. [16] employed KL-divergence to mine clustering patterns from concatenated complementary information, using consistent semantics to constrain the alignments of representation distributions in each view [16]. Zhou and Shen [17] applied inter-view weight and game learning to aggregate complementary and consistent information, and then resorted to structure differences between data and the simplex to optimize the model for mining patterns of cross-view data of dances [17]. Gao et al. [4] designed dual invariant semantic learning for fusing information of views, and then modeled the reinforcement clustering process to capture an adaptive partition policy for mining patterns of cross-view data of dances [4]. Xu et al. [18] proposed a fusion-free cross-view auto encoder architecture that captured cluster complementarity between representations of cross-view data of images in a high-dimensional non-linear mapping manner. This approach aimed to align clustering structures hidden in heterogeneous manifolds of views and utilized an E-like optimization between information aggregations and structure explorations to enhance pattern mining [18]. DealMVC designed a dual contrastive cross-view clustering network by aligning inter-view and intra-view feature-level similarity graphs and class-level pseudo-label graphs, then enforced intra-view pseudo-label graphs to be as similar as possible to obtain clustering results [19]. DuaLGR developed a cross-view graph refinement clustering method through mutual guidance between complementary semantic fusion and dual label explorations to enhance consensus graph refinement between views [20]. HCLS_CGL designed a confidence neighbor cross-view graph clustering method that captured group-wise structures between data in fusing consistent and complementary information to drive consensus graph learning between views [21].

3. Deep mutual information decoupling based unsupervised image clustering

Consider a cross-view image dataset $X = \{X^1, X^2, \ldots, X^V\}$ with n samples and V views, $X^v = \{x_i^v\}, |$ is the v-th view set, and x_i^v is the i-th data in the v-th view. A novel mutual information decoupling based generative cross-view representation learning is proposed to partition cross-view images X into K groups in an unsupervised manner, which is comprised of information maximization deep generative module and self-supervised posterior inference module, as illustrated in Fig. 1.

3.1. Information Maximization Deep Generative Module

DMID-UIC models the joint probability distribution between data and clusters to generate cross-view data with

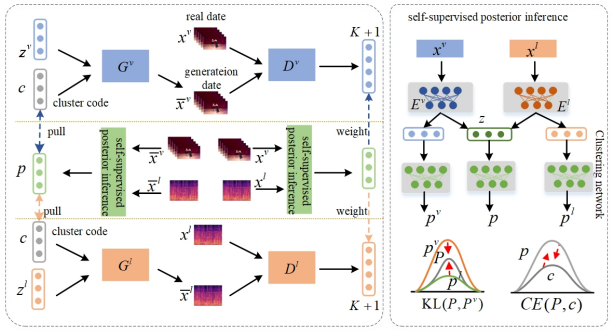

Fig. 1. The illustration of DMID-UIC, which is comprised of information maximization deep generative module and self-supervised posterior inference module

the help of generative adversarial network with class priors, and then maximizes the mutual information between generation data and clusters to decouple clustering features and redundant features in cross-view data, which effectively enhances semantic robustness of models.

Specifically, due to the diversity of cross-view data, DMID-UIC deploys specific generative adversarial networks for each view to fully fit data distributions. Taking the v-th view as an example, in the generative stage, DMID-UIC utilizes the continuous noise code z and the discrete cluster code c as the input of the generator G^v, to generate data X^v that belong to specific classes, which is expressed as follows:

$$\bar{X}^v = G^v(z, c) \tag{1}$$

where c is subject to the polynomial distribution $P(c;\phi)$ with the parameter ϕ. DMID-UIC utilizes a mixture of discrete and continuous latent variables to create a non-smooth geometric latent space, which effectively promotes the separation of different features in the data space and prevents the interference from overlapping feature distributions on clusters. In the adversarial stage, DMID-UIC utilizes the discriminator D^v to distinguish real data X^v and generated data $\bar{X}^v$, which is expressed as follows:

$$D^v\left(X^v\right) \in R^{n(K+1)} \tag{2}$$

$$D^v\left(\bar{X}^v\right) \in R^{n(K+1)} \tag{3}$$

where K denotes numbers of clusters. Compared to a typical discriminator that makes binary determinations regarding data authenticity by defining a mapping from the data space to real values, DMID-UIC attaches discriminators with $K+1$ output units, which increases the judgment of the classes of data on the basis of data true or false judgments, thereby providing better guidance for enhancing the quality of generated data. Then, a novel objective function is designed to achieve the aforementioned generative adversarial learning,

$$L = L_G + L_D \tag{4}$$

where the losses of the generators L_G and the discriminators L_D are as follows,

$$L_G = -\frac{1}{2}\sum_{v=1}^{V} \mathrm{E}_{x^v\sim P_f} \exp\left\{\sigma^{-1}\left[D_i^v\left(X^v\right)\right]\right\} \tag{5}$$

$$\begin{aligned} L_D = \sum_{v=1}^{V}\Bigg\{ & \mathrm{E}_{x^v\sim P_f}\sum_{i=1}^{K+1}\left[\log\left(1-D_i^v\left(X^v\right)\right)\right] \\ & +\mathrm{E}_{x_{\text{real}}^v/P_r}\sum_{i=1}^{K+1}\left[w_i\log\left(D_i^v\left(x_{\text{real}}^v\right)\right)\right]\Bigg\} \end{aligned} \tag{6}$$

where $\sigma = \frac{1}{1+e^{-x}}$ is the logistic sigmoid. $w_i = P\left(c=i \mid \left\{X_{\text{real}}^{(v)}\right\}\right)$ is the soft assignment probability of the i-th class generated by the self-supervised posterior inference module for data, which serves as the weight of each output unit of the discriminators. D_i^v represents the value of the i-th output unit of in the v-th view discriminator.

To fully capture the common clustering information shared by views and prevent the loss of cluster information during the generation process. DMID-UIC implements mandatory decoupling of cluster information and redundant information via maximizing mutual information between cluster code and generated data.

$$I\left(c,\bar{X}^v\right) = \sum_{v=1}^{V} I\left(c, G^v(z,c)\right) = \sum_{v=1}^{V}\left(H(c) - H\left(c \mid G^v(z,c)\right)\right) \tag{7}$$

where $I\left(c,\bar{X}^v\right)$ denotes the mutual information between cluster code and generated data. $H(c)$ is the Shannon entropy of the cluster code. $H\left(c \mid G^v(z,c)\right)$ is the conditional entropy.

Since the mutual information computation requires estimating the posterior distribution, it is challenging to compute. Therefore, DMID-UIC utilizes variational information maximization to provide an approximate variational lower bound for the original optimization problem. The variational lower bound can be expressed as:

$$\begin{aligned} I\left(c,G^v(z,c)\right) &= H(c) - H\left(c \mid G^v(z,c)\right) \\ &= \mathrm{E}_{x^v\sim G^v(z,c)}\left[\mathrm{E}_{c\sim P(c|x^v)}\left[\log P\left(c \mid x^v\right)\right]\right] + H(c) \\ &= \mathrm{E}_{x^v\sim G^v(z,c)}\left[\mathrm{KL}\left(P\left(c \mid x^v\right) \| Q\left(c \mid x^v\right)\right)\right. \\ &\quad \left. + \mathrm{E}_{c\sim P(c|x^v)}\left[\log Q\left(c \mid x^v\right)\right]\right] + H(c) \\ &\geq = \mathrm{E}_{x^v\sim G^v(z,c)}\mathrm{E}_{c\sim P(c|x^v)}\left[\log Q\left(c \mid x^v\right)\right] + H(c) \end{aligned} \tag{8}$$

where $Q\left(c \mid x^v\right)$ is an approximate distribution of a posterior probability $P\left(c \mid x^v\right)$. To optimize $Q\left(c \mid x^\nu\right)$, DMID-UIC designs a self-supervised posterior inference module which minimizes the difference between the posterior distribution of the generated data and the prior of the common cluster code, achieving approximation to the true posterior.

3.2. Self-supervised Posterior Inference Module

In this module, DMID-UIC implements the fitting of the posterior distribution for generated data and the cluster information of cross-view data in the latent space.

Specifically, DMID-UIC utilizes view-specific encoder networks to map real data of each view into common latent spaces:

$$Z^v = E^v\left(X^v\right) \tag{9}$$

where $E^v(\cdot)$ denotes the v-th view encoder. Z^v denotes latent representations of X^v. Then, DMID-UIC fuses consistent and complementary information of each view via utilizing latent representations $\left[Z^1, Z^2, \ldots, Z^V\right]$ to construct the common representation Z, i.e., $Z = F\left(\left[Z^v\right]_{v=1}^V\right)$. $F(\cdot)$ is the fusion function, e.g., the linear concatenation.

After obtaining common representations of cross-view data, DMID-UIC constructs a clustering network with Student's distribution as the core to soft assignment probability, as follows:

Specifically, a learnable neural network is introduced to generate global pseudo labels of Z, i.e., $f(Z;Q): Z \rightarrow P$, where $Q = \{q_1, q_2, \ldots, q_k\}$ stands for pattern prototypes of Z and P stands for membership distribution of clusters:

$$p = C(Z;Q) = \zeta(\varphi(Z,Q)) \tag{10}$$

$\varphi(\cdot)$ denote structure differences between date and pattern prototypes:

$$p_{ij} = \varphi\left(z_i, q_j\right) = \frac{1}{1 + \left\|z_i - q_j\right\|} \tag{11}$$

where p_{ij} denotes the similarity between the i-th representation and the j-th prototype. $\zeta(\cdot)$ maps structure differences to the cluster membership:

$$p_{ij} = \zeta\left(p_i\right) = \frac{\left(p_{ij} / \sum_j p_{ij}\right)^2}{\sum_j \left(p_{ij} / \sum_j p_{ij}\right)^2} \tag{12}$$

After training the model, global pseudo labels of data are gained,

$$y_i = \arg\max\left(p_{ij}\right) \tag{13}$$

Then, the cross entropy between global pseudo labels and cluster code is utilized to enhance the semantic discriminability of the cluster code:

$$L_{ce} = -\sum_{i=1}^{n}\sum_{j=1}^{k} p_{ij} \log c_{ij} \tag{14}$$

Furthermore, the Kullback-Leibler divergence between P of Z and p^v of z^v is utilized to improve robustness of global pseudo labels:

$$L_{con} = -\sum_{v=1}^{V}\sum_{i=1}^{n}\sum_{j=1}^{k} p_{ij} \log \frac{p_{ij}}{p_{ij}^v} \tag{15}$$

where p^v is obtained via inputting z^v into the clustering network.

In addition, to provide discriminative guidance from real data to the information maximization deep generative module and enhance feature decoupling capabilities, in each round of the alternating training process, the self-supervised posterior inference module weight soft assignments P of real data to output units of the discriminators. This approach achieves mutual guidance and joint optimization between the two modules, which effectively enhances the robustness of DMID-UIC.

3.3. Model Optimization

The overall loss function L of DMID-UIC is:

$$L = \min_{G,E}\max_{D} L(G,D,E) = L_G + \alpha L_D - \lambda\left(L_{ce} + L_{con}\right) \tag{16}$$

where λ and α are trade-off parameters.

DMID-UIC uses the expectation-maximization (EM) algorithm to handle hidden variables like cluster codes. This involves iterative optimization through the E-step (expectation calculation) and M-step (parameter maximization) to solve the maximum likelihood function.

For the v-th view data, the maximum likelihood function is:

$$L(\theta) = \log P\left(X^v \mid \theta\right) \tag{17}$$

where θ is the optimization parameter set. In the E-step, we calculate the posterior probability distribution of the hidden variables (variables that are not directly observed) based on the current estimates of the parameters. The goal here is to estimate the likely values of these hidden variables given the observed data and the current set of parameters. Essentially, this step involves making educated guesses about the unseen parts of the data using the knowledge we already have. The E-step calculates the posterior probability distribution of the hidden variable c:

$$\tilde{P}^{(i+1)}(c) = P\left(c \mid X^v, \theta^{(i)}\right) \tag{18}$$

where $\theta^{(i)}$ is the optimal parameter from the previous M-step. In the M-step, we search for a new set of parameters that maximize the expected log-likelihood of the data, given the posterior distribution of the hidden variables calculated in the E-step. This means finding the parameters that best explain the observed data under the current assumptions. After this step, we obtain a new set of parameters that will be used in the next iteration of the E-step. The M-step finds the optimal parameter $\theta^{(i+1)}$:

$$L(\theta) = F\left(\tilde{P}^{i+1}, \theta^{(i+1)}\right) \tag{19}$$

Table 1. The average clustering results on Fashion dataset

Metric	DealMVC	COMIC	DULGR	HCLS_CGL	VCGA	DMCAG	CoMVC	DMID-UIC
ACC	0.7243	0.5561	0.5815	0.5825	0.7653	0.8039	0.8358	**0.9212**
NMI	0.7683	0.5991	0.6243	0.6241	0.8069	0.8467	0.8772	**0.8917**
PUR	0.6942	0.6143	0.6673	0.7102	0.7766	0.8186	0.8459	**0.9424**

Table 2. The average clustering results on MNIST-USPS dataset

Metric	DealMVC	COMIC	DULGR	HCLS_CGL	VCGA	DMCAG	CoMVC	DMID-UIC
ACC	0.7463	0.4603	0.6751	0.6990	0.9561	0.9598	0.9664	**0.9892**
NMI	0.6536	0.6870	0.7245	0.7841	0.9427	0.9409	0.9499	**0.9724**
PUR	0.7470	0.5101	0.6807	0.7579	0.9518	0.9598	0.9660	**0.9892**

Table 3. The average clustering results on MNIST-USPS dataset

Metric	DealMVC	COMIC	DULGR	HCLS_CGL	VCGA	DMCAG	CoMVC	DMID-UIC
ACC	0.4405	0.4008	0.4939	0.3938	0.4913	0.4872	0.4452	**0.5923**
NMI	0.3271	0.4251	0.4589	0.2352	0.4490	0.4502	0.4051	**0.5198**
PUR	0.4748	0.5137	0.5427	0.4060	0.5134	0.5361	0.5063	**0.5832**

Table 4. The average clustering results on Unige-Maastricht dataset

Metric	DealMVC	COMIC	DULGR	HCLS_CGL	VCGA	DMCAG	CoMVC	DMID-UIC
ACC	0.6106	0.7361	0.6798	0.8419	0.7238	0.8302	0.8681	**0.8922**
NMI	0.6536	0.7791	0.7228	0.8849	0.7668	0.8732	0.8711	**0.8885**
PUR	0.7167	0.7027	0.6933	0.8888	0.7258	0.7469	0.8936	**0.9267**

The optimal parameter θ^{i+1} is:

$$\theta^{(i+1)} = \arg\max_{\theta} Q\left(\theta, \theta^{(i)}\right) \tag{20}$$

In this context, the discriminator and generator (often used in Generative Adversarial Networks) are adjusted based on their performance in distinguishing real data from generated data. The discriminator's goal is to better differentiate between real and generated data, while the generator aims to produce data that is as realistic as possible, thereby fooling the discriminator. The generator's loss function is defined as:

$$L_G = -\frac{1}{2}\sum_{v=1}^{V} \mathrm{E}_{x^v \sim P_f} \exp\left\{\sigma^{-1}\left[D_i^v\left(x^v\right)\right]\right\} \tag{21}$$

The discriminator's loss function is defined as:

$$\begin{aligned} L_D = \sum_{v=1}^{V} \Bigg\{ & \mathrm{E}_{x^v \sim P_f} \sum_{i=1}^{K+1} \left[\log\left(1 - D_i^v\left(x^v\right)\right)\right] \\ & + \mathrm{E}_{x_{\text{real}}^v \sim P_r} \sum_{i=1}^{K+1} \left[w_i \log\left(D_i^v\left(x_{\text{real}}^v\right)\right)\right] \Bigg\} \end{aligned} \tag{22}$$

Through continuous iterations of the E-step and M-step, the model repeatedly updates its parameters, gradually reducing the gap between the generated data and the real data. The final goal is to find a set of optimal parameters that minimize the difference between the generated and real data, thereby maximizing the likelihood of the data.

4. Result

4.1. Setup

Dataset: Four common datasets in cross-view image clustering are used to prove the effectiveness and superiority of DMID-UIC in recognizing image patterns. The statistical information is displayed in the following. Fashion is a dataset of images of products with 10000 samples of 10 classes where three images of same class are considered as three views of samples. MNIST-USPS is a dataset of the handwritten digit with 5000 samples of 10 classes where MNIST and USPS are usually considered as two views of samples. Caltech101-7 is a dataset of the object detection with 1474 samples of 7 classes where Wavelet Moments features and Census Transform Histogram features are selected as two views of samples. Unige-Maastricht is a dataset of videos with 1474 samples of 20 classes where Mel-scale Frequency Cepstral Coefficients, Scale-invariant feature transform features, and Space-Time Interest Point

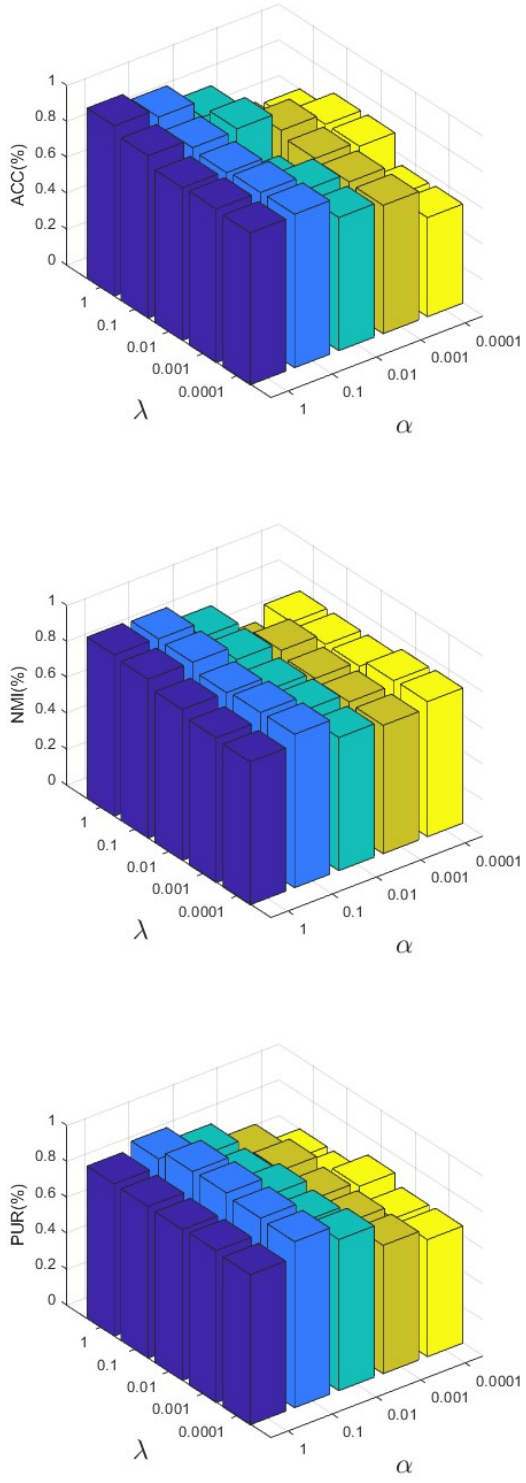

Fig. 2. The sensitivity analysis of parameters α and λ on Fashion

features are selected as three views of dances.

Evaluation Metric: Three clustering metrics are used to quantitatively evaluate the performance of DMID-UIC, i.e., Accuracy (ACC), Normalized Mutual Information (NMI), and Purity (PUR). Greater values for the three metrics signify better clustering performance. The experiments undergo ten repetitions, and the final results is determined by averaging ten results to ensure a fair comparison.

Comparison Methods: The seven cross-view clustering methods are used as comparison baselines, i.e., DealMVC, COMIC, DuaLGR, HCLS_CGL, VCGA, DMCAG, and CoMVC.

Implementation details: DMID-UIC is implemented by PyTorch. In the experiments, vector representations of samples are normalized to $[0, 1]$, and then are fed into the network for training. In the training, the Adam solver is used with the batchsize 100 to optimize the network, and the learning rates are set to $[0.001, 0.0001]$ for all modules of the network on four datasets. The trade-off parameter λ is set as 0.1 on four datasets.

4.2. Comparisons with State-of-the-arts

The comparison results with State-of-the-arts are illustrated in Tables 1 to 4, and it can be noticed that DMID-UIC achieves the best clustering performance on four datasets regarding ACC, NMI, and PUR, which establishes a fresh state-of-the-art benchmark for mining patterns of dances. Specifically, DMID-UIC is about 8.49% and 7.89% higher than the second-best comparison results in term of ACC and PUR on the Fashion dataset, respectively. The reasons for the performance improvement of DMID-UIC are twofold. (1) DMID-UIC maximizes the mutual information between data and clusters to achieve decoupling between clustering features and redundant features hidden in multi-view data, which captures a clustering structure with clear boundaries. (2) DMID-UIC utilizes EM optimization strategy to promote the information maximization deep generative learning and the self-supervised posterior inference learning in an iterative manner, which effectively models the joint distribution of multi-view data and clusters. In addition, the clustering performance of deep clustering methods is superior to that of deep fusion methods. The reason is that deep clustering methods can utilizes the cluster information of data assignments to improve aggregation of complementary information of multi-view data compared to two-stage learning in deep fusion methods.

4.3. Parameter analysis

Parameter Analysis is conducted in terms of α and λ on Fashion. Specifical α and λ are constrained in the set $\{0.0001, 0.001, 0.01, 0.1, 1\}$. The results displayed in Fig. 2

show performance changes of DMID-UIC with α and λ which reveals insightful trends regarding the behavior of DMID-UIC with respect to changes in α and λ Notably, when both α and λ are set to 0.1, DMID-UIC achieves the best performance across all settings. The optimal performance at $\alpha = \lambda = 0.1$ indicates that the model's components are equally sensitive to the adjustments made during the training process, and this balance is critical for the model to learn and generalize well.

5. Conclusions

We propose a novel generative cross-view clustering method, named DMID-UIC, aimed at recognizing image patterns. DMID-UIC leverages mutual information decoupling to enhance pattern recognition by effectively modeling the joint probability distribution between data and clusters. By maximizing mutual information, DMID-UIC disentangles clustering features from redundant features, leading to improved clarity in capturing clustering structures with distinct boundaries. The method incorporates an expectation-maximization (EM) optimization strategy to iteratively refine the learning process. This involves alternating between information maximization deep generative learning and self-supervised posterior inference learning. Through this iterative refinement, DMID-UIC effectively fits the joint distribution of multi-view data and clusters, addressing both feature extraction and clustering challenges. Extensive qualitative and quantitative experimental results demonstrate that DMID-UIC sets a new state-of-the-art benchmark in image pattern recognition, highlighting its effectiveness in achieving accurate and meaningful clustering outcomes.

References

[1] K. Cui, R. Li, S. L. Polk, Y. Lin, H. Zhang, J. M. Murphy, R. J. Plemmons, and R. H. Chan, (2024) *"Superpixel-based and Spatially-regularized Diffusion Learning for Unsupervised Hyperspectral Image Clustering"* **IEEE Transactions on Geoscience and Remote Sensing**: DOI: 10.1109/TGRS.2024.3385202.

[2] H. Huang, C. Wang, X. Wei, and Y. Zhou, (2024) *"Deep Image Clustering: A survey"* **Neurocomputing**: 128101. DOI: 10.1016/j.neucom.2024.128101.

[3] P. Li, J. Gao, J. Zhang, S. Jin, and Z. Chen, (2022) *"Deep Reinforcement Clustering"* **IEEE Transactions on Multimedia *25***: 8183–8193. DOI: 10.1109/TMM.2022.3233249.

[4] J. Gao, M. Liu, P. Li, J. Zhang, and Z. Chen, (2023) *"Deep Multiview Adaptive Clustering With Semantic Invariance"* **IEEE Transactions on Neural Networks and Learning Systems**: DOI: 10.1109/TMM.2022.3233249.

[5] P. Li, Z. Chen, L. T. Yang, J. Gao, Q. Zhang, and M. J. Deen, (2018) *"An incremental deep convolutional computation model for feature learning on industrial big data"* **IEEE Transactions on Industrial Informatics *15***(3): 1341–1349. DOI: 10.1109/TII.2018.2871084.

[6] C. Cui, Y. Ren, J. Pu, J. Li, X. Pu, T. Wu, Y. Shi, and L. He, (2024) *"A novel approach for effective multi-view clustering with information-theoretic perspective"* **Advances in Neural Information Processing Systems *36***:

[7] Q. Xiao, J. Dai, J. Luo, and H. Fujita, (2019) *"Multi-view manifold regularized learning-based method for prioritizing candidate disease miRNAs"* **Knowledge-Based Systems *175***: 118–129. DOI: 10.1016/j.knosys.2019.03.023.

[8] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. "Deep canonical correlation analysis". In: *International conference on machine learning*. 2013, 1247–1255.

[9] C. Zhang, Y. Liu, and H. Fu. "Ae2-nets: Autoencoder in autoencoder networks". In: *computer vision and pattern recognition*. 2019, 2577–2585. DOI: 10.1109/CVPR.2019.00268.

[10] Z. Wan, C. Zhang, P. Zhu, and Q. Hu. "Multi-view information-bottleneck representation learning". In: *AAAI conference on artificial intelligence*. ***35***. 11. 2021, 10085–10092. DOI: 10.1609/aaai.v35i11.17210.

[11] Q. Gao, H. Lian, Q. Wang, and G. Sun. "Cross-modal subspace clustering via deep canonical correlation analysis". In: *AAAI Conference on artificial intelligence*. ***34***. 04. 2020, 3938–3945. DOI: 10.1609/aaai.v34i04.5808.

[12] M. Yang, Y. Li, P. Hu, J. Bai, J. Lv, and X. Peng, (2022) *"Robust multi-view clustering with incomplete information"* **IEEE Transactions on Pattern Analysis and Machine Intelligence *45***(1): 1055–1069. DOI: 10.1109/TPAMI.2022.3155499.

[13] C. Cui, Y. Ren, J. Pu, X. Pu, and L. He. "Deep multi-view subspace clustering with anchor graph". In: *International Joint Conference on Artificial Intelligence*. 2023, 3577–3585. DOI: 10.48550/arXiv.2305.06939.

[14] Z. Gu and S. Feng, (2023) *"Individuality meets commonality: A unified graph learning framework for multi-view clustering"* **ACM Transactions on Knowledge Discovery from Data *17***(1): 1–21. DOI: 10.1145/3532612.

[15] D. J. Trosten, S. Lokse, R. Jenssen, and M. Kampffmeyer. "Reconsidering representation alignment for multi-view clustering". In: *computer vision and pattern recognition*. 2021, 1255–1265. DOI: 10.1109/CVPR46437.2021.0013.

[16] J. Xu, Y. Ren, H. Tang, Z. Yang, L. Pan, Y. Yang, X. Pu, S. Y. Philip, and L. He, (2022) *"Self-supervised discriminative feature learning for deep multi-view clustering"* **IEEE Transactions on Knowledge and Data Engineering *35***(7): 7470–7482. DOI: 10.1109/TKDE.2022.3193569.

[17] R. Zhou and Y.-D. Shen. "End-to-end adversarial-attention network for multi-modal clustering". In: *computer vision and pattern recognition*. 2020, 14619–14628.

[18] J. Xu, C. Li, Y. Ren, L. Peng, Y. Mo, X. Shi, and X. Zhu. "Deep incomplete multi-view clustering via mining cluster complementarity". In: *conference on artificial intelligence*. ***36***. 8. 2022, 8761–8769. DOI: 10.1609/aaai.v36i8.20856.

[19] X. Yang, J. Jiaqi, S. Wang, K. Liang, Y. Liu, Y. Wen, S. Liu, S. Zhou, X. Liu, and E. Zhu. "Dealmvc: Dual contrastive calibration for multi-view clustering". In: *ACM International Conference on Multimedia*. 2023, 337–346. DOI: 10.1145/3581783.361195.

[20] Y. Ling, J. Chen, Y. Ren, X. Pu, J. Xu, X. Zhu, and L. He. "Dual label-guided graph refinement for multi-view graph clustering". In: *AAAI Conference on Artificial Intelligence*. ***37***. 7. 2023, 8791–8798. DOI: 10.1609/aaai.v37i7.26057.

[21] J. Wen, C. Liu, G. Xu, Z. Wu, C. Huang, L. Fei, and Y. Xu. "Highly confident local structure based consensus graph learning for incomplete multi-view clustering". In: *Computer Vision and Pattern Recognition*. 2023, 15712–15721.