# Cross-view Locality Preserved Diversity and Consensus Learning for Multi-view Unsupervised Feature Selection

Chang Tang, *Member, IEEE*, Xiao Zheng, Xinwang Liu, *Senior Member, IEEE*, Wei Zhang, *Member, IEEE*, Jing Zhang, Jian Xiong, *Member, IEEE*, and Lizhe Wang, *Fellow, IEEE*

**Abstract**—Although demonstrating great success, previous multi-view unsupervised feature selection (MV-UFS) methods often construct a view-specific similarity graph and characterize the local structure of data within each single view. In such a way, the cross-view information could be ignored. In addition, they usually assume that different feature views are projected from a latent feature space while the diversity of different views cannot be fully captured. In this work, we resent a MV-UFS model via cross-view local structure preserved diversity and consensus learning, referred to as CvLP-DCL briefly. In order to exploit both the shared and distinguishing information across different views, we project each view into a label space, which consists of a consensus part and a view-specific part. Therefore, we regularize the fact that different views represent same samples. Meanwhile, a cross-view similarity graph learning term with matrix-induced regularization is embedded to preserve the local structure of data in the label space. By imposing the $l_{2,1}$-norm on the feature projection matrices for constraining row sparsity, discriminative features can be selected from different views. An efficient algorithm is designed to solve the resultant optimization problem and extensive experiments on six publicly datasets are conducted to validate the effectiveness of the proposed CvLP-DCL.

**Index Terms**—Multi-view unsupervised feature selection, local structure preservation, feature projection, diversity and consensus learning, cross-view similarity graph.

## 1 Introduction

With the rapid development of data acquisition sensors and data processing technologies, data are often represented by different feature descriptors. For an instance, in image/video processing, different visual descriptors such as Scale Invariant Feature Transform (SIFT) [1] , Local Binary Patterns (LBP) [2], and Histogram of Oriented Gradient (HOG) [3] are often used to describe each image/video frame from different views. In biomedical research, for different cells, the chemical response as well as the chemical structure can be used to represent a certain drug, while the sequence and gene expression values can represent a certain protein in different aspects [4], [5]. In general, the data represented by multiple views are regarded as multi-view data in data mining and machine learning communities. In the last decades, a variety of multi-view learning techniques have been put forward to process the multi-view data [6], [7], [8], [9], [10], [11], [12]. As a special case, multi-view unsupervised feature selection (MV-UFS), which promotes many learning task by selecting a small feature subset from original multi-view data, has obtained more and more attention since different views of data are usually with high dimensionality and processing these data is confronted with the curse of dimensionality problem [13]. In addition, it is a challenging and time-consuming task to obtain the labels from large number of data instances.

In the past few years, a variety of MV-UFS methods have been introduced and they can be mainly categorized into two classes. The first class of approaches first combines different feature views together and then uses traditional single-view UFS methods such as Laplacian score [14], trace ratio [15], spectral feature selection [16] and minimum redundancy spectral feature selection [17] are carried out on the concentrated data. This kind of methods cannot exploit the underlying correlations between different views. Instead of concentrating different views, the other class of MV-UFS methods aim to build models from multi-view data directly, and they often excavate the diversity and complementary information to promote the feature selection performance. Typical methods in this class include Adaptive Multi-View Feature Selection (AMFS) method [18], Adaptive Unsupervised Multi-View Feature Selection (AUMFS) [19], Adaptive Similarity and View Weight (ASVW) learning for Multi-View Feature Selection [20], Robust Multi-View Feature Selection (RMFS) [21] and Consensus Learning Guided Multi-view Unsupervised Feature Selection (CGMV-UFS) [22]. Since the diversity and complementary information are important for multi-view learning, MV-UFS methods

- *C. Tang and L. Wang are with the School of Computer Science, China University of Geosciences, Wuhan 430074, P.R. China (E-mail: tangchang@cug.edu.cn; lizhe.wang@gmail.com).*
- *X. Zheng and X. Liu are with the School of Computer, National University of Defense Technology, Changsha 410073, P.R. China (E-mail: zxnudt@gmail.com, xinwangliu@nudt.edu.cn).*
- *W. Zhang is with Shandong Provincial Key Laboratory of Computer Networks, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250000, P.R. China (E-mail: wzhang@qlu.edu.cn).*
- *J. Zhang is with the College of Software, Beihang University, Beijing 100191, P.R. China (E-mail: zhang_jing@buaa.edu.cn).*
- *J. Xiong is with School of Business Administration, Southwestern University of Finance and Economics, Chengdu, Sichuan, 611130, China (E-mail: xiongjian2017@swufe.edu.cn).*

*Manuscript received xx xx, xxxx; revised August xx, xxxx. (Corresponding authors: Xinwang Liu and Wei Zhang)*

in the second class often perform better than those in the first category. In this work, we focus on the second class of methods which our proposed approach belongs to.

Without label information, the local properties of samples usually act as a priori to serve the feature selection task. Therefore, traditional methods usually use various similarity graphs to characterize the local geometrical manifold structure of data and then rank the importance of each feature. However, previous approaches often construct a view-specific similarity graph within each single view, while the cross view information could be ignored. In addition, in order to capture the shared structure of different views, a certain consensus latent feature space is often learned and different feature views are assumed to be generated from this space while the effect of the diversity and noises of different views on the projection has not been taken into account. In order to address above two issues, we present a cross-view locality preserved diversity and consensus learning model for MV-UFS, referred to as CvLP-DCL briefly. Instead of projecting multiple views into a consensus latent feature space, we project each view of original data into a label space. In order to capture both the shared and distinguishing information of different views, the label apace is relaxed to a consensus part and a diversity part. In such a way, different feature views are regularized to represent the same samples. Different to previous multi-view learning methods which exploit the common information and specificity of different views in the original feature space, we capture these properties in the label space, which is more straightforward and reasonable since multiple views must share the same sample labels. Meanwhile, instead of using only a view-specific similarity graph to preserve the local structure of different samples in a single view separately, we integrate a cross-view similarity graph learning term with matrix-induced regularization into the model.

This manuscript is an extension of the AAAI conference version [23], and it differs [23] with following significant additional contributions:

- Instead of directly combining the inter-view similarity graph between pairwise views and the intra-view similarity graph in each single view, we design a cross-view similarity graph learning term with matrix-induced regularization to learn a collaborative similarity graph from each single graph for preserving the locality of data for MV-UFS;
- By using the matrix-induced regularization, the importance of different views can be adaptively learned for serving the cross-view similarity graph learning;
- More extensive experimental comparisons are conducted to evaluate and analyze the proposed method.

## 2 Related Work

In this part, we briefly introduce some recent research works about MV-UFS. In [18], Wang et al. proposed a MV-UFS method for human motion retrieval, in which the multi-view of local feature descriptors are used to represent human motion data. For each view of data, a graph Laplacian matrix is generated, then these view-specific Laplacian matrices are linearly combined with weights to exploit complementary information of different views. Finally, trace ratio criteria is deployed to eliminate redundant features. In order to identify important feature dimensions, AUMFS [19] uses an $l_{2,1}$-norm regularized sparse regression model to project original data into cluster labels. In AUMFS, the $l_{2,1}$-norm is used to impose row sparsity on the projection matrix for measuring feature importance. In addition, the local geometrical structure of data is preserved by the linearly combined weighted view-specific graph Laplacian matrices. In RMFS [21], robust multi-view k-means is used to obtain the pseudo labels for sparse feature selection, the pseudo labels are generated by utilizing the heterogeneous information from multiple views. By considering that previous methods such as AMFS and AUMFS ignore the underlying shared structure across different feature views, and the pre-computed similarity matrices are not accurate for characterizing the local structure of data, ASVW [20] leverages the learning mechanism to adaptively learn a similarity graph shared by different views. To further learn a compact feature representation, Wan et al. [24] proposed to reduce original high-dimensional data to low dimensions and unified different views to a combination weight matrix. In order to capture both the common and complementary information of different views, CGMV-UFS [22] constructs a view-dependent graph Laplacian matrix for each view for intra-view local structure preservation. Meanwhile, CGMV-UFS learns a common label indicator matrix to regularize that different feature views represent the same samples. However, as aforementioned, almost all of previous methods are confronted with at least two issues, i.e., the cross-view local structure is not taken into consideration and the assumption of projecting multi-view data into a single label space is too strict since there are noises and specificity in each single view.

## 3 Proposed Method

### 3.1 Notations

Throughout this paper, matrices and vectors are denoted as boldface capital letters and boldface lower case letters, respectively. For an arbitrary matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, $M_{ij}$ denotes its $(i,j)$-th entry, $\mathbf{m}^i$ and $\mathbf{m}_j$ denote its $i$-th row and $j$-th column, respectively. $\mathrm{Tr}(\mathbf{M})$ is the trace of $\mathbf{M}$ if $\mathbf{M}$ is square and $\mathbf{M}^\top$ is the transpose of $\mathbf{M}$. $\mathbf{I}_m$ is the identity matrix with size $m \times m$ (denoted by $\mathbf{I}$ if the size is obviously known). The $l_{2,1}$-norm of matrix $\mathbf{M}$ is defined as $||\mathbf{M}||_{2,1} = \sum_{i=1}^{m} ||\mathbf{m}^i|| = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} {M_{ij}}^2}$. $||\mathbf{M}||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} {M_{ij}}^2}$ is the well-known Frobenius norm of $\mathbf{M}$. $||\mathbf{M}||_1 = \sum_{i=1}^{m}\sum_{j=1}^{n} |M_{ij}|$ represents the $l_1$-norm of matrix $\mathbf{M}$, i.e., the absolute summation of its entries.

Supposing there are $N$ data samples $\{\mathbf{x}_i\}_{i=1}^{N}$ which belong to $c$ classes, and they are characterized by $V$ different views of features, the data matrix is denoted as $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{d\times N}$. Let $\mathbf{x}_i^v$ denote the $v$-th view of the $i$-th sample, then the complete $i$-th sample $\mathbf{x}_i = [\mathbf{x}_i^1; \cdots; \mathbf{x}_i^V] \in \mathbb{R}^d$ consists of features from $V$ views, and the dimension of the $v$-th view $\mathbf{x}_i^v \in \mathbb{R}^{d_v}$ is $d_v$, such that $d = \sum_{v=1}^{V} d_v$. The data matrix of the $v$-th view can be represented as $\mathbf{X}^v = [\mathbf{x}_i^v, \cdots, \mathbf{x}_N^v] \in \mathbb{R}^{d_v\times N}$, then

$\mathbf{X} = [\mathbf{X}^v; \cdots; \mathbf{X}^v]$. The task of MV-UFS is to select the top $K$ discriminative features from those $d$ features without label information of data instances.

### 3.2 Formulation of CvLP-DCL

Although a lot of data consist of multi-view heterogeneous features, they still share the same semantic information. In order to capture this common information, we project different views of features into a shared label space, which represent original data in a relatively higher level manner. Considering that each single view contains both the common information and distinguishing specificity, we relax the common label space to a consensus part and a diversity part, this can be mathematically formulated as follows:

$$\min_{\mathbf{W}^v, \bar{\mathbf{Y}}, \mathbf{Y}^v} \sum_{v=1}^{V} \mathcal{L}(\mathbf{X}^v, \mathbf{W}^v, \bar{\mathbf{Y}}, \mathbf{Y}^v) + \xi \sum_{v=1}^{V} \mathcal{R}(\mathbf{W}^v, \mathbf{Y}^v), \quad (1)$$

where $\mathcal{L}(\mathbf{X}^v, \mathbf{W}^v, \bar{\mathbf{Y}}, \mathbf{Y}^v)$ is the projection operator for the $v$-th view, and $\mathcal{R}(\mathbf{W}^v, \mathbf{Y}^v)$ denotes certain regularization on $\mathbf{W}^v$ and $\mathbf{Y}^v$. $\xi$ is a positive constant for balancing the two terms. $\mathbf{W}^v \in \mathbb{R}^{d_v \times c}$ is the projection matrix for the $v$-th view, $\bar{\mathbf{Y}} \in \mathbb{R}^{N \times c}$ and $\mathbf{Y}^v \in \mathbb{R}^{N \times c}$ represents the consensus part and the diversity part of the label space, respectively. Since $\bar{\mathbf{Y}}$ denotes the pure label indicator matrix of data, we constrain it as $\bar{\mathbf{Y}} \in \{0,1\}^{N \times c}$. However, it is difficult to solve Eq. (1) with this discrete constrain, we use the orthogonality constraint instead, i.e., $\bar{\mathbf{Y}}^\top \bar{\mathbf{Y}} = \mathbf{I}, \bar{\mathbf{Y}} \geq 0$. In this work, we use the regression model to formulate the projection process, which can be written as:

$$\mathcal{L}(\mathbf{X}^v, \mathbf{W}^v, \bar{\mathbf{Y}}, \mathbf{Y}^v) = ||\mathbf{X}^{v\top}\mathbf{W}^v - (\bar{\mathbf{Y}} + \mathbf{Y}^v)||_F^2. \quad (2)$$

In Eq. (2), the projected label space is decomposed into a consensus part for capturing the consensus label representation of different views and a diversity part for capturing the distinct diversity of each view.

In order to select discriminative features, we impose row sparsity on $\mathbf{W}^v$ by using the $l_{2,1}$-norm regularization. In addition, although each view contains some view-specific information, they still represent the same data, and thus the consensus label representation should be the main part. For $\mathbf{Y}^v$, it just denotes the distinct variance or noisy information for the $v$-th view. Therefore, we do not impose the orthogonality constraint on $\mathbf{Y}^v$, but impose the $l_1$-norm instead to constrain its sparsity. Finally, $\mathcal{R}(\mathbf{W}^v, \mathbf{Y}^v)$ can be formulated as:

$$\mathcal{R}(\mathbf{W}^v, \mathbf{Y}^v) = ||\mathbf{W}^v||_{2,1} + ||\mathbf{Y}^v||_1. \quad (3)$$

For unsupervised feature selection [25], [26], [27], [28], [29], the local geometrical structure works as a crucial priori. In our work, we also preserve the local geometrical structure of data by learning a collaborative similarity graph from different views. Given $V$ similarity graphs constructed from $V$ different views $\{\mathbf{S}^v\}_{v=1}^V$, we formulate the collaborative similarity graph learning as follows:

$$\min_{\bar{\mathbf{S}}, \boldsymbol{\gamma}} \|\bar{\mathbf{S}} - \sum_{v=1}^{V} \gamma^v \mathbf{S}^v\|_F^2, s.t. \gamma^v > 0, \forall v, \bar{\mathbf{S}}\mathbf{1} = \mathbf{1}, \bar{S}_{ij} \geq 0, \quad (4)$$

where $\bar{\mathbf{S}}$ is the collaborative similarity graph which needs to be learned, $\boldsymbol{\gamma} = [\gamma^1, \gamma^2, \cdots, \gamma^V] \in \mathbb{R}^{V \times 1}$ is a vector which consists of the view weights of different views, $\mathbf{S}^v$ is the sample similarity matrix and each of its entry is defined as $S_{ij}^v = \exp(-||\mathbf{x}_i^v - \mathbf{x}_j^v||^2/\sigma^2)$ .

By using the learned collaborative similarity graph, we regularize that if two data samples are similar to each other, their learned label vectors should also be close to each other, of which the regularization can be derived as:

$$\min_{\bar{\mathbf{S}}, \bar{\mathbf{Y}}} \sum_{i,j=1}^{N} ||\bar{\mathbf{y}}^i - \bar{\mathbf{y}}^j||_2^2 \bar{S}_{ij} = \min_{\bar{\mathbf{S}}, \bar{\mathbf{Y}}} \mathrm{Tr}(\bar{\mathbf{Y}}^\top \mathbf{L}_{\bar{\mathbf{S}}} \bar{\mathbf{Y}}), \quad (5)$$

where $\mathbf{L}_{\bar{\mathbf{S}}} = \bar{\mathbf{D}} - \bar{\mathbf{S}}$ is the cross-view Laplacian matrix and $\bar{\mathbf{D}}$ is a diagonal matrix with its $i$-th diagonal entry calculated as the sum of the $i$-th row in $\bar{\mathbf{S}}$, i.e., $\bar{\mathbf{D}}_{ii} = \sum_{j=1}^{N} \bar{S}_{ij}$.

In order to prevent two highly similar views from being allocated with large weights simultaneously, we add a matrix-induced regularization term to learn diverse weights for different views, and the model can be mathematically formulated as:

$$\min_{\boldsymbol{\gamma} \in \mathbb{R}_+^{V \times 1}} \sum_{p,q=1}^{V} \gamma^p \gamma^q M_{pq} = \boldsymbol{\gamma}^\top \mathbf{M} \boldsymbol{\gamma}, \quad s.t. \ \boldsymbol{\gamma}^\top \mathbf{1} = \mathbf{1}, \quad (6)$$

where $\mathbf{M}$ is a matrix which measures the correlation between different views. For two similarity matrices $\mathbf{S}^p$ and $\mathbf{S}^q$, if they are highly similar to each other, the corresponding row/columns of two matrices will be highly related, and the inner product of the row/column vectors from two matrices should be large. Therefore, the sum of all the inner product values, i.e., $\mathrm{Tr}(\mathbf{S}^p \mathbf{S}^q)$ should also be large. As a result, we define $M_{pq} = \mathrm{Tr}(\mathbf{S}^{p\top} \mathbf{S}^q)$.

By combining Eq.(2)-Eq.(6) together, we obtain the mathematical optimization model of CvLP-DCL as follows:

$$\begin{aligned} \min_{\mathbf{W}^v, \bar{\mathbf{Y}}, \mathbf{Y}^v, \bar{\mathbf{S}}, \boldsymbol{\gamma}} & \sum_{v=1}^{V} \Big\{ \|(\mathbf{X}^{v\top})\mathbf{W}^v - (\bar{\mathbf{Y}} + \mathbf{Y}^v)\|_F^2 \\ & + \gamma^v \|\mathbf{Y}^v\|_1 + (1-\gamma^v)\|\mathbf{W}^v\|_{2,1}\Big\} \\ & + \lambda \mathrm{Tr}(\bar{\mathbf{Y}}^\top \mathbf{L}_{\bar{\mathbf{S}}} \bar{\mathbf{Y}}) + \alpha \|\bar{\mathbf{S}} - \sum_{v=1}^{V} \gamma^v \mathbf{S}^v\|_F^2 + \frac{\beta}{2} \boldsymbol{\gamma}^\top \mathbf{M} \boldsymbol{\gamma} \\ s.t. \ & \bar{\mathbf{Y}}^\top \bar{\mathbf{Y}} = \mathbf{I}, \bar{\mathbf{Y}} \geq 0. M_{pq} = \mathrm{Tr}(\mathbf{S}^{p\top} \mathbf{S}^q), \\ & \boldsymbol{\gamma}^\top \mathbf{1} = \mathbf{1}, \gamma^v > 0, \forall v. \ \bar{\mathbf{S}}\mathbf{1} = \mathbf{1}, \bar{S}_{ij} \geq 0. \end{aligned} \quad (7)$$

As can be seen from Eq.(7), we use the learned view weights to regularize $\|\mathbf{Y}^v\|_1$ and $\|\mathbf{W}^v\|_{2,1}$ instead of fixing their hyper parameters. There are two advantages: On one hand, if $\gamma^v$ is large, it means that the $v$-th view is important for clustering, i.e., the $v$-th view is more likely to be close to the consensus of different views. Therefore, the specificity of the $v$-th view (represented as $\mathbf{Y}^v$) should not be obvious. On the other hand, if the $v$-th view is important, its features should also be allocated with large weights during the feature selection process, i.e., $||\mathbf{W}^v||_{2,1}$ should be large. Therefore, we put $\gamma^v$ and $1-\gamma^v$ on $\mathbf{Y}^v$ and $\mathbf{W}^v$ respectively to regularize the diversity distribution and feature selection capability. Moreover, the local geometrical structure of data samples is preserved in the label space via the learned collaborative similarity graph Laplacian regularization term. In our CvLP-DCL model, the label learning,

feature selection and collaborative similarity graph learning are integrated into a unified framework, in which different learning tasks can promote each other to obtain there final optimized solutions. Note that the work in [30] also learns a consensus guidance for unsupervised feature selection, our proposed model is different to [30] in following aspects: 1) Although the work in [30] also maps original features into label space for multi-view feature selection, it dose not take the diversity information of different views into account during the mapping process. In our proposed CvLP-DCL, we decompose the mapped label space into two parts: a consensus part which is shared by all of the views and a diversity part which is unique for each single view; 2) In [30], different basic partitions must be given in advance, while CvLP-DCL automatically generates label matrix during the learning process; 3) In CvLP-DCL, we propose a collaborative similarity graph learning model from different views, the learned similarity graph is used to preserve the local geometrical structure of original data in the label space; 4) In addition, we use a matrix-induced regularization term to learn diverse weights for different feature views, which can prevent two highly similar views from being allocated with large weights simultaneously.

### 3.3 Optimal Solution of CvLP-DCL

The variables need to be solved in Eq.(7) include the projection matrices $\mathbf{W}^v$, label matrices $\bar{\mathbf{Y}}$ and $\mathbf{Y}^v$, collaborative similarity matrix $\bar{\mathbf{S}}$, and the vector $\boldsymbol{\gamma}$ which consists of view weights. Since these variables are related to each other, it is difficult to solve them at one step. Hence, we design an alternative iterative algorithm to solve the optimization problem. At each time, we optimize the objective function w.r.t one variable with others fixed and the procedure repeats until meeting the convergence condition.

#### 3.3.1 Optimize $\bar{\mathbf{Y}}$ by Fixing Other Variables

When other variables are fixed, $\bar{\mathbf{Y}}$ can be obtained by solving the following problem:

$$\begin{aligned}\min_{\bar{\mathbf{Y}}} \sum_{v=1}^{V} \|\mathbf{X}^{v\top}\mathbf{W}^v - (\bar{\mathbf{Y}} - \mathbf{Y}^v)\|_F^2 + \lambda \mathrm{Tr}(\bar{\mathbf{Y}}^\top \mathbf{L}_s \bar{\mathbf{Y}}) \\ s.t.\ \bar{\mathbf{Y}}^\top\bar{\mathbf{Y}} = \mathbf{I}, \bar{\mathbf{Y}} \geq 0.\end{aligned} \tag{8}$$

Then, Eq.(8) can be rewritten as the following equal trace form:

$$\begin{aligned}&\min_{\bar{\mathbf{Y}}} \sum_{v=1}^{V} \mathrm{Tr}(-2(\mathbf{W}^v)^\top \mathbf{X}^v \bar{\mathbf{Y}} + 2(\mathbf{Y}^v)^\top \bar{\mathbf{Y}}) + \mathrm{Tr}(\bar{\mathbf{Y}}^\top\bar{\mathbf{Y}}) \\ &+ \lambda \mathrm{Tr}(\bar{\mathbf{Y}}^\top \mathbf{L}_{\bar{\mathbf{S}}} \bar{\mathbf{Y}}), \\ &s.t.\ \ \bar{\mathbf{Y}}^\top\bar{\mathbf{Y}} = \mathbf{I}, \bar{\mathbf{Y}} \geq 0.\end{aligned} \tag{9}$$

By adding an extra penalty term $\rho||\bar{\mathbf{Y}}^\top\bar{\mathbf{Y}} - \mathbf{I}||_F^2$ and introduce a Lagrange multiplier $\Phi$ to eliminate the orthogonal constraint and remove the inequality constraint, respectively. Then we have the following Lagrange function:

$$\begin{aligned}\mathcal{F}(\bar{\mathbf{Y}}, \Phi) = &\sum_{v=1}^{V} \mathrm{Tr}(-2(\mathbf{W}^v)^\top \mathbf{X}^v \bar{\mathbf{Y}} + 2(\mathbf{Y}^v)^\top \bar{\mathbf{Y}}) + \mathrm{Tr}(\bar{\mathbf{Y}}^\top\bar{\mathbf{Y}}) \\ &+ \lambda \mathrm{Tr}(\bar{\mathbf{Y}}^\top \mathbf{L}_{\bar{\mathbf{S}}} \bar{\mathbf{Y}}) + \rho||\bar{\mathbf{Y}}^\top\bar{\mathbf{Y}} - \mathbf{I}||_F^2 - \mathrm{Tr}(\Phi^\top\bar{\mathbf{Y}}).\end{aligned} \tag{10}$$

By taking the derivative of $\mathcal{F}(\bar{\mathbf{Y}}, \Phi)$ w.r.t $\bar{\mathbf{Y}}$, and setting it to zero, we have

$$\begin{aligned}\frac{\partial F(\bar{\mathbf{Y}}, \Phi)}{\partial \bar{\mathbf{Y}}} = &\sum_{v=1}^{V} 2\mathbf{Y}^v - 2(\mathbf{X}^v)^\top \mathbf{W}^v + 2\lambda \mathbf{L}_{\bar{\mathbf{S}}} \bar{\mathbf{Y}} \\ &+ 2\bar{\mathbf{Y}} + 4\rho\bar{\mathbf{Y}}(\bar{\mathbf{Y}}^\top\bar{\mathbf{Y}} - \mathbf{I}) - \Phi = 0.\end{aligned} \tag{11}$$

Then, we can get $\Phi$:

$$\Phi = 2(\mathbf{I} + \lambda \mathbf{L}_{\bar{\mathbf{S}}})\bar{\mathbf{Y}} + 4\rho\bar{\mathbf{Y}}(\bar{\mathbf{Y}}^\top\bar{\mathbf{Y}} - \mathbf{I}) + \sum_{v=1}^{V} 2\mathbf{Y}^v - 2(\mathbf{X}^v)^\top \mathbf{W}^v \tag{12}$$

Based on the Karush-Kuhn-Tucker condition [31], i.e., $\Phi_{ij}\bar{Y}_{ij} = 0$, we get the following equation:

$$[2(\mathbf{I} + \lambda \mathbf{L}_{\bar{\mathbf{S}}})\bar{\mathbf{Y}} + 4\rho\bar{\mathbf{Y}}(\bar{\mathbf{Y}}^\top\bar{\mathbf{Y}} - \mathbf{I}) + \sum_{v=1}^{V} 2\mathbf{Y}^v - 2\mathbf{X}^{v\top}\mathbf{W}^v]_{ij}\bar{Y}_{ij} = 0. \tag{13}$$

Then, $\bar{\mathbf{Y}}$ can be updated via following strategy:

$$\bar{Y}_{ij} \leftarrow \bar{\mathbf{Y}}_{ij} \frac{[2\rho\bar{\mathbf{Y}} + \sum_{v=1}^{V} \mathbf{X}^{v\top}\mathbf{W}^v]_{ij}}{[(\mathbf{I} + \lambda \mathbf{L}_{\bar{\mathbf{S}}})\bar{\mathbf{Y}} + 2\rho\bar{\mathbf{Y}}\bar{\mathbf{Y}}^\top\bar{\mathbf{Y}} + \sum_{v=1}^{V} \mathbf{Y}^v]_{ij}}. \tag{14}$$

In this work, in order to constrain the orthogonality of $\bar{\mathbf{Y}}$, we set $\rho$ a relatively large value, $\rho = 10^6$ in our experiments.

#### 3.3.2 Optimize $\mathbf{Y}^v$ by Fixing Other Variables

By fixing other variables, $\mathbf{Y}^v$ can be updated by solving the following optimization problem:

$$\min_{\mathbf{Y}^v} \ \|\mathbf{X}^{v\top}\mathbf{W}^v - (\bar{\mathbf{Y}} - \mathbf{Y}^v)\|_F^2 + \gamma^v\|\mathbf{Y}^v\|_1, \tag{15}$$

which can be solved by using the soft-thresholding operator obtained as follows:

$$\mathbf{Y}^v = sign(\mathbf{X}^{v\top}\mathbf{W}^v - \bar{\mathbf{Y}})max(|\mathbf{X}^{v\top}\mathbf{W}^v - \bar{\mathbf{Y}}| - \frac{\gamma^v}{2}, \mathbf{0}). \tag{16}$$

#### 3.3.3 Optimize $\mathbf{W}^v$ with Other Variables Fixed

When other variables are fixed, solving $\mathbf{W}^v$ is equal to the following problem:

$$\min_{\mathbf{W}^v} \ \|\mathbf{X}^{v\top}\mathbf{W}^v - (\bar{\mathbf{Y}} - \mathbf{Y}^v)\|_F^2 + (1 - \gamma^v)\|\mathbf{W}^v\|_{2,1}, \tag{17}$$

Which can be solved by using the iterative re-weighted least-squares algorithm. Since there is an $l_{2,1}$-norm regularization in Eq.(17), we cannot obtain its closed form solution. By taking the derivative of objective function in Eq.(17) w.r.t $\mathbf{W}^v$ and setting it to zero, we obtain:

$$\mathbf{X}^v\mathbf{X}^{v\top}\mathbf{W}^v - \mathbf{X}^v(\bar{\mathbf{Y}} + \mathbf{Y}^v) + (1 - \gamma^v)\mathbf{G}^v\mathbf{W}^v = 0 \tag{18}$$

where $\mathbf{G}^v$ is a diagonal matrix with its $i$-th diagonal entry calculated as $\mathbf{G}^v_{ii} = \frac{1}{2\|\mathbf{W}^v(i,:)\|_2^2}$. According to Eq.(18), $\mathbf{W}^v$ can be updated as:

$$\mathbf{W}^v = (\mathbf{X}^v\mathbf{X}^{v\top} + (1 - \gamma^v)\mathbf{G}^v)^{-1}(\mathbf{X}^v(\bar{\mathbf{Y}} + \mathbf{Y}^v). \tag{19}$$

Then $\mathbf{W}^v$ and $\mathbf{G}^v$ can be updated in an alternative manner.

#### 3.3.4 Optimize $\bar{\mathbf{S}}$ by Fixing Other Variables

With other variables fixed, we can obtain $\bar{\mathbf{S}}$ by solving the following problem:

$$\begin{aligned} &\min_{\bar{\mathbf{S}}} \lambda \mathrm{Tr}(\bar{\mathbf{Y}}^\top \mathbf{L}_{\bar{\mathbf{S}}} \bar{\mathbf{Y}}) + \alpha \|\bar{\mathbf{S}} - \sum_{v=1}^{V} \gamma^v \mathbf{S}^v\|_F^2 \\ &s.t.\ \bar{\mathbf{S}}\mathbf{1} = \mathbf{1}, \bar{S}_{ij} \geq 0. \end{aligned} \tag{20}$$

By setting $\Theta_{ij} = ||\bar{\mathbf{y}}^i - \bar{\mathbf{y}}^j||_2^2$, and $\mathbf{S}_{\boldsymbol{\gamma}} = \sum_{v=1}^{V} \gamma^v \mathbf{S}^v$, Eq.(20) can be rewritten as:

$$\begin{aligned} &\min_{\bar{\mathbf{S}}} \alpha \|\bar{\mathbf{S}} - \mathbf{S}_\gamma\|_F^2 + \frac{1}{2}\lambda \sum_{i,j=1}^{N} \Theta_{ij} \bar{S}_{ij} \\ &s.t.\ \bar{\mathbf{S}}\mathbf{1} = \mathbf{1}, \bar{S}_{ij} \geq 0. \end{aligned} \tag{21}$$

Eq.(21) can be solved by using alternative optimization of each row of $\bar{\mathbf{S}}$. For the $i$-th row, the corresponding optimization problem is equal as follows:

$$\begin{aligned} &\min_{\mathbf{s}_i} \alpha \|\bar{\mathbf{s}}_i - (\mathbf{s}_{\gamma i} - \frac{\lambda}{4\alpha}\boldsymbol{\theta}_i)\|_F^2 \\ &s.t.\ \bar{\mathbf{s}}_i \mathbf{1} = 1, \bar{\mathbf{s}}_i \geq 0. \end{aligned} \tag{22}$$

Let $\boldsymbol{\eta} = \bar{\mathbf{s}}_i \top$ and $\boldsymbol{\nu} = (\mathbf{s}_{\gamma i} - \frac{\lambda}{4\alpha}\boldsymbol{\theta}_i)\top$, then the Lagrangian function of problem (22) can be written as:

$$\min_{\boldsymbol{\eta}} \ \|\boldsymbol{\eta} - \boldsymbol{\nu}\|_2^2 - \tau(\boldsymbol{\eta}^\top \mathbf{1} - 1) - \boldsymbol{\xi}^\top \boldsymbol{\eta} \tag{23}$$

where $\tau$ and $\boldsymbol{\xi}$ are a scalar and a Lagrangian coefficient vector, respectively. Denoting the optimal solution and the associate Lagrangian coefficients of above problem as $\boldsymbol{\eta}^*$ , $\tau^*$ and $\boldsymbol{\xi}^*$. According to the KKT condition [31], we can establish the following equations:

$$\begin{cases} \forall i, & \eta_i^* - \nu_j - \tau^* - \xi_i^* = 0, \\ \forall i, & \eta_i^* \geq 0, \\ \forall i, & \xi_i^* \geq 0, \\ \forall i, & \eta_i^* \xi_i^* = 0, \end{cases} \tag{24}$$

where $\eta_i^*$ is the $i$-th element of $\boldsymbol{\eta}^*$. Based on the constraint $\boldsymbol{\eta}^\top \mathbf{1} = 1$, we have $\tau^* = \frac{1 - \mathbf{1}^\top \boldsymbol{\nu} - \mathbf{1}^\top \boldsymbol{\xi}^*}{N}$ and $\boldsymbol{\eta}^* = (\boldsymbol{\nu} - \frac{\mathbf{1}\mathbf{1}^\top}{N}\boldsymbol{\nu} + \frac{1}{N}\mathbf{1} - \frac{\mathbf{1}^\top \boldsymbol{\xi}^*}{N}\mathbf{1}) + \boldsymbol{\xi}^*$.

Let $\hat{\xi}^* = \frac{\mathbf{1}^\top \boldsymbol{\xi}^*}{N}$ and $\mathbf{u} = \boldsymbol{\nu} - \frac{\mathbf{1}\mathbf{1}^\top}{N}\boldsymbol{\nu} + \frac{1}{N}\mathbf{1}$, then $\boldsymbol{\eta}^* = \mathbf{u} + \boldsymbol{\xi}^* - \hat{\xi}^* \mathbf{1}$. So $\forall i$ we have

$$\eta_i^* = u_i + \xi_i^* - \hat{\xi}^* \tag{25}$$

According to Eqs.(24) and (25) we know $u_i + \xi_i^* - \hat{\xi}^* = (u_i - \hat{\xi}^*)_+$, here $x_+ = max(x, 0)$. Then we have

$$\eta_i^* = (u_i - \hat{\xi}^*)_+ \tag{26}$$

Therefore, if $\hat{\boldsymbol{\xi}}^*$ is known, the optimal solution $\boldsymbol{\eta}^*$ can be easily obtained.

Note that Eq.(25) can be rewritten as $\xi_i^* = \eta_i^* + \hat{\xi}^* - u_i$. Similarly, we have $\xi_j^* = (\hat{\xi}^* - u_j)_+$ and $\hat{\boldsymbol{\xi}}^* = \frac{1}{N-1}\sum_{i=1}^{N-1}(\hat{\xi}^* - u_i)_+$ based on Eq.(24). Defining a function as

$$f(\hat{\xi}) = \frac{1}{N-1}\sum_{i=1}^{N-1}(\hat{\xi}^* - u_i)_+ - \hat{\xi}. \tag{27}$$

Let $f(\hat{\xi}) = 0$ and we can solve the root finding problem to obtain $\xi^*$.

Note that $\xi^* \geq 0$, $f'(\hat{\xi})$ and $f'(\hat{\xi})$ is convex and piecewise linear, the Newton method is used to find the root of $f(\hat{\xi}) = 0$ efficiently, i.e,

$$\hat{\xi}_{t+1} = \hat{\xi}_t - \frac{f(\hat{\xi})}{f'(\hat{\xi})}. \tag{28}$$

In such a manner, $\boldsymbol{\eta}^*$ , $\tau^*$ and $\boldsymbol{\xi}^*$ can be optimized alternatively.

#### 3.3.5 Optimize $\boldsymbol{\gamma}$ by Fixing Other Variables

When other variables are fixed, the view weights can be learned by solving following problem:

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \ & \gamma^v \|\mathbf{F}^v\|_1 + (1 - \gamma^v)\|\mathbf{W}^v\|_{2,1} + \alpha \|\bar{\mathbf{S}} - \sum_{v=1}^{V} \gamma^v \mathbf{S}^v\|_F^2 \\ & + \frac{\beta}{2}\boldsymbol{\gamma}^\top \mathbf{M}\boldsymbol{\gamma} \\ & s.t.\ \boldsymbol{\gamma}^\top \mathbf{1} = \ \mathbf{1}, \gamma^v > 0, \forall v \end{aligned} \tag{29}$$

Eq.(29) can be transferred to a quadratic programming problem with linear constraints as follows:

$$\begin{aligned} &\min_{\boldsymbol{\gamma}} \ \frac{2\alpha + \beta}{2}\boldsymbol{\gamma}^\top \mathbf{M}\boldsymbol{\gamma} + \mathbf{f}^\top \boldsymbol{\gamma} \\ &s.t.\ \boldsymbol{\gamma}^\top \mathbf{1} = 1, \gamma^v > 0, \forall \gamma \end{aligned} \tag{30}$$

where $M_{ij} = \mathrm{Tr}(\mathbf{S}^{i\top}\mathbf{S}^j)$, $\mathbf{f}_i = \|\mathbf{F}^v\|_1 + \|\mathbf{W}^v\|_{2,1} - 2\alpha \mathrm{Tr}(\mathbf{S}^\top \mathbf{S}^i)$.

Eq.(30) can be easily solved by using some standard quadratic programming solving toolbox.

We summarize the optimization procedure of CvLP-DCL in Algorithm 1.

**Algorithm 1** Iterative algorithm for solving CvLP-DCL

**Input:** Multi-view data matrices $\{\mathbf{X}^v \in \mathbb{R}^{d_v \times N}\}_{v=1}^V$, parameters: $\lambda, \alpha$, and $\beta$. A small constant $\varepsilon = 0.0000001$.

**Initialize:** $\mathbf{Y}^1, \cdots, \mathbf{Y}^v, \mathbf{W}^1, \cdots, \mathbf{W}^v$. $\bar{\mathbf{S}} = \frac{1}{N}\sum_{v=1}^{V} \mathbf{S}^v$

**while not converged do**

1. Update $\bar{\mathbf{Y}}$ by solving Eq.(8);
2. Update $\mathbf{Y}^v$ via Eq.(16);
3. Update $\mathbf{W}^v$ by solving Eq.(17);
4. Update $\bar{\mathbf{S}}$ by solving Eq.(20);
5. Update $\boldsymbol{\gamma}$ by solving Eq.(30);
6. Check convergence condition: $(obj^{t-1} - obj^t)/obj^t < \varepsilon$.

**end while**

**Output:** $\mathbf{W}^1, \cdots, \mathbf{W}^v$.

**Feature selection:** Sort the $l_2$-norm of the rows of $\{\mathbf{W}^v\}_{v=1}^V$ in decent order and select the largest $K$ values. The feature dimension indexes with the the largest $K$ values are selected to form the feature subset.

### 3.4 Theoretical Analysis of the Proposed Algorithm

In this section, we give a brief theoretical analysis of Algorithm 1, including convergence analysis and complexity analysis.

### 3.5 Convergence Analysis

Since there are a number of variables need to be alternatively updated in Algorithm 1, it is not easy to give a detailed theoretical convergence proof. Here we give a brief convergence analysis of each single step. In step 1 of Algorithm 1, since we use the Karush-Kuhn-Tucker condition to update $\bar{\mathbf{Y}}$, the objective value of Eq.(9) can be ensured to monotonically decrease. In step 2 for updating $\mathbf{Y}^v$, the soft-thresholding operator can ensure the global optimal solution of Eq.(15). As to step 3, $\mathbf{W}^v$ and $\mathbf{G}^v$ are iteratively updated via the iterative re-weighted least-squares algorithm, of which the convergence can be also guaranteed. For updating $\bar{\mathbf{S}}$ in step 4, the Karush-Kuhn-Tucker condition and the Newton method can also ensure the convergence. As to the final step for updating $\boldsymbol{\gamma}$, the convergence of the quadratic programming problem can be also guaranteed. In the experimental section, the strong convergence behaviour of the proposed algorithm will also be empirically validated in the experiments section.

### 3.6 Time Complexity Analysis

For updating $\bar{\mathbf{Y}}$, the main computation lies in calculating Eq.(14), which only consists of some matrix multiplication operations, and its computational complexity is $\mathcal{O}(N\max(d_v, N)c)$. As to updating $\mathbf{Y}^v$, there also only consists of a matrix multiplication operation, i.e., $(\mathbf{X}^v)^\top \mathbf{W}^v$, of which the computational complexity is $\mathcal{O}(Nd_vc)$. For solving $\mathbf{W}^v$, since we need to compute the inverse of a $d_v \times d_v$ matrix, the computational complexity is $\mathcal{O}(d_v^3)$. As to solving $\bar{\mathbf{S}}$ and $\boldsymbol{\gamma}$, they have linear computational complexity. Therefore, the total main computation complexity of Algorithm 1 is $\mathcal{O}(N\max(d_v, N)c + T_1(d_v^3))$ for each iteration, where $T_1$ is the inner iteration time for updating $\mathbf{W}^v$.

## 4 Experiments

In this section, extensive experiments are conducted to evaluate the performance of CvLP-DCL on some real-world benchmark datasets. In addition, we compare the proposed CvLP-DCL with several other state-of-the-art UFS methods to validate its efficacy.

### 4.1 Datasets

In this work, six publicly available multi-view datasets are used in our experiments. They are:

**Handwritten** is obtained from the UCI machine learning repository [32], and consists of handwritten digits from 0 to 9. There are 2000 data samples in total and each sample is described by 6 different features.

**Caltech101-7** is an image dataset captured for object recognition problem [33]. There are 101 different categories of images in this dataset. Following previous works [34], [35], 7 classes with 1474 images are used in our experiments. Six different features are extracted for each image.

**Reuters** is a documents dataset which consists of five different languages and their translations [36]. There are 6 classes of all the documents. We use the subset that are written in English and all their translations in all the other 4 languages (French, German, Spanish and Italian).

**NUSWIDEOBJ** is a dataset for object recognition. There are 30000 images in 31 categories [37] in total. Five different features for each sample are used in our experiments.

**MSRCV1** is an image dataset which contains 240 images with 8 object classes [38]. Seven classes of samples including tree, building, airplane, cow, face, car and bicycle are selected in our experiments, and each sample is represented by 6 types of features.

**BBCSport** is a documents dataset collected from the BBC Sport website, the content corresponds to 5 topical areas of sports news, two kinds of feature are entreated for each sample. [39].

The detailed information including feature dimensions and types of these datasets are summarized in Table 1.

### 4.2 Experimental Setup

Similar to previous works [20], [22], [24], [40], the $k$-means clustering is performed on the selected features to evaluate the performance of our proposed CvLP-DCL. Two commonly used evaluation metrics including accuracy (ACC) and normalized mutual information (NMI) are used to evaluate the quality of the selected feature subsets obtained from different feature selection algorithms. Let $t_i$ and $r_i$ respectively represent the true label of $x_i$ and the clustering results. Then ACC can be defined as follows:

$$ACC = \frac{\sum_{i=1}^{N} \delta(r_i, map(t_i))}{N}, \tag{31}$$

where $\delta(x, y) = 1$ if $x = y$, otherwise $\delta(x, y) = 0$. $map(t_i)$ is the best mapping function which permutes clustering labels to match the true labels using the KuhnMunkres algorithm. Given two variables $R$ and $T$, NMI is defined as

$$NMI(R, T) = \frac{I(R, T)}{\sqrt{H(R)H(T)}}, \tag{32}$$

where $H(R)$ and $H(T)$ are the entropies of $R$ and $T$, respectively, and $I(R, T)$ is the mutual information between $R$ and $T$. For clustering task, $R$ and $T$ are the clustering results and the true labels, respectively. NMI reflects the degree of correlation between clustering results and ground truth labels. For both two metrics, larger values represent better performance.

Meanwhile, We also compare the proposed CvLP-DCL with other seven different single view and multi-view unsupervised feature selection methods, they are as follows:

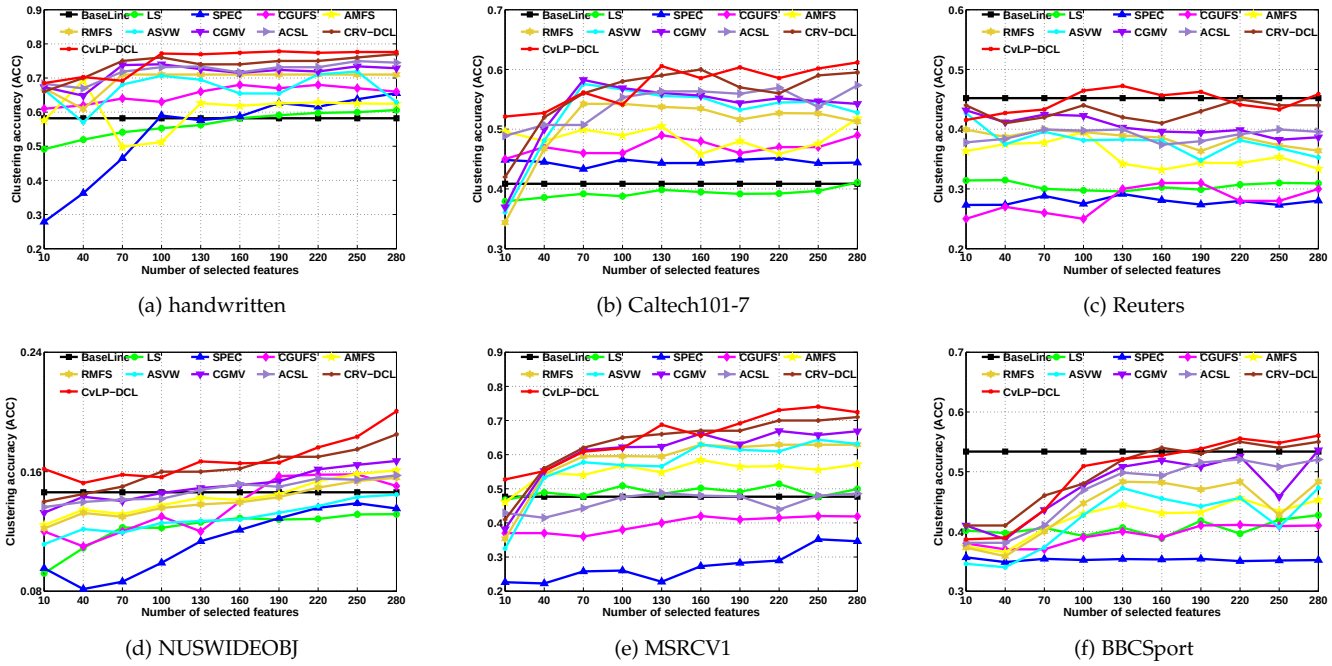
- **Baseline**: As the most classical and basic clustering algorithm, $k$-means is used for clustering by simply combining all features into a single view.
- Laplacian score (**LS**) [14] and spectral feature selection (**SPEC**) [16]: Two representative and classical single view unsupervised feature selection methods. The multi-view features are also firstly combined together for these two algorithms;
- Consensus guided unsupervised feature selection (**CGUFS**) [30], which is a single-view unsupervised feature selection method and introduces consensus clustering to generate pseudo labels for feature selection. We also combine multi-view features together for this algorithm;
- **AMFS** [18]: Adaptive multi-view feature selection, which is an unsupervised feature selection approach

TABLE 1
The detailed information of the multi-view datasets used in our experiments

| Feature index | Handwritten | Caltech101-7 | Reuters | NUSWIDEOBJ | MSRCV1 | BBCSport |
|---|---|---|---|---|---|---|
| 1 | Pix(240) | Gabor(48) | English(21531) | CH(65) | CENT(1302) | View one(3183) |
| 2 | Fou(76) | WM(40) | France(24892) | CM(226) | CMT(48) | View two(3203) |
| 3 | Fac(216) | CENTRIST(254) | German(34251) | CORR(145) | GIST(512) | - |
| 4 | ZER(47) | HOG(1984) | Italian(15506 | EDH(74) | HOG(100) | - |
| 5 | KAR(64) | GIST(512) | Spanish(11547) | WT(129) | LBP(256) | - |
| 6 | MOR(6) | LBP(928) | - | - | SIFT(210) | - |
| No. of samples | 2000 | 1474 | 18758 | 26315 | 210 | 544 |
| No. of classes | 20 | 7 | 6 | 31 | 7 | 5 |

TABLE 2
Clustering results (ACC% $\pm$ std%) of different algorithms on different datasets by implementing $K$-means and GMC on the selected features.

| Datasets | handwritten | | Caltech101-7 | | Reuters | | NUSWIDEOBJ | | MSRCV1 | | BBCSport | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | K-means | 58.20±4.89 | K-means | 40.86±3.70 | K-means | 45.20±2.51 | K-means | 14.62±0.43 | K-means | 47.67±2.87 | K-means | 53.37±1.41 |
| | GMC | 65.34 | GMC | 43.92 | GMC | 47.34 | GMC | 16.56 | GMC | 48.93 | GMC | 56.75 |
| LS | K-means | 60.71±5.32 | K-means | 41.17±3.37 | K-means | 31.42±1.01 | K-means | 13.26±0.31 | K-means | 52.21±5.65 | K-means | 43.04±4.25 |
| | GMC | 64.23 | GMC | 44.85 | GMC | 34.45 | GMC | 15.27 | GMC | 54.82 | GMC | 47.72 |
| SPEC | K-means | 65.53±6.47 | K-means | 45.15±2.67 | K-means | 27.20±0.00 | K-means | 14.06±0.46 | K-means | 36.74±5.41 | K-means | 36.05±0.10 |
| | GMC | 66.74 | GMC | 49.02 | GMC | 28.52 | GMC | 16.35 | GMC | 39.04 | GMC | 40.85 |
| CGUFS | K-means | 68±4.89 | K-means | 49.31±3.08 | K-means | 31.79±0.98 | K-means | 15.82±0.37 | K-means | 42.31±4.27 | K-means | 41.06±1.22 |
| | GMC | 71.54 | GMC | 53.66 | GMC | 36.15 | GMC | 17.04 | GMC | 48.92 | GMC | 47.84 |
| AMFS | K-means | 69.41±1.81 | K-means | 52.37±2.86 | K-means | 39.84±1.31 | K-means | 16.10±0.38 | K-means | 58.41±4.96 | K-means | 48.02±1.12 |
| | GMC | 72.92 | GMC | 55.77 | GMC | 43.73 | GMC | 18.53 | GMC | 59.62 | GMC | 52.74 |
| RMFS | K-means | 71.04±3.21 | K-means | 54.37±2.64 | K-means | 39.94±1.24 | K-means | 16.23±0.53 | K-means | 62.94±5.27 | K-means | 48.32±1.07 |
| | GMC | 73.48 | GMC | 58.07 | GMC | 44.65 | GMC | 19.16 | GMC | 65.67 | GMC | 53.27 |
| ASVW | K-means | 72.13±4.91 | K-means | 56.24±5.18 | K-means | 41.48±1.97 | K-means | 16.52±0.49 | K-means | 65.41±4.62 | K-means | 51.77±1.21 |
| | GMC | 73.91 | GMC | 57.92 | GMC | 45.82 | GMC | 20.08 | GMC | 67.79 | GMC | 56.86 |
| CGMV-UFS | K-means | 75.45±5.99 | K-means | 58.25±5.46 | K-means | 43.16±2.33 | K-means | 17.25±0.40 | K-means | 68.93±6.22 | K-means | 54.03±1.05 |
| | GMC | 77.84 | GMC | 60.63 | GMC | 47.88 | GMC | 21.52 | GMC | 70.81 | GMC | 59.57 |
| ACSL | K-means | 73.28±4.05 | K-means | 57.68±4.20 | K-means | 40.47±2.62 | K-means | 15.57±0.46 | K-means | 49.38±5.33 | K-means | 52.34±1.14 |
| | GMC | 75.66 | GMC | 60.4 | GMC | 45.39 | GMC | 19.45 | GMC | 52.39 | GMC | 58.35 |
| CRV-DCL | K-means | 76.47±4.23 | K-means | 59.23±5.25 | K-means | 45.07±2.14 | K-means | 17.86±0.39 | K-means | 69.58±5.72 | K-means | 54.95±1.16 |
| | GMC | 79.06 | GMC | 62.58 | GMC | 49.36 | GMC | 22.06 | GMC | 71.85 | GMC | 60.17 |
| CvLP-DCL | K-means | 77.89±4.04 | K-means | 61.06±4.17 | K-means | 46.78±2.27 | K-means | 20.03±0.63 | K-means | 72.22±5.31 | K-means | 56.68±1.32 |
| | GMC | 82.71 | GMC | 64.61 | GMC | 49.84 | GMC | 22.85 | GMC | 75.79 | GMC | 61.53 |




(a) handwritten (b) Caltech101-7 (c) Reuters

(d) NUSWIDEOBJ (e) MSRCV1 (f) BBCSport

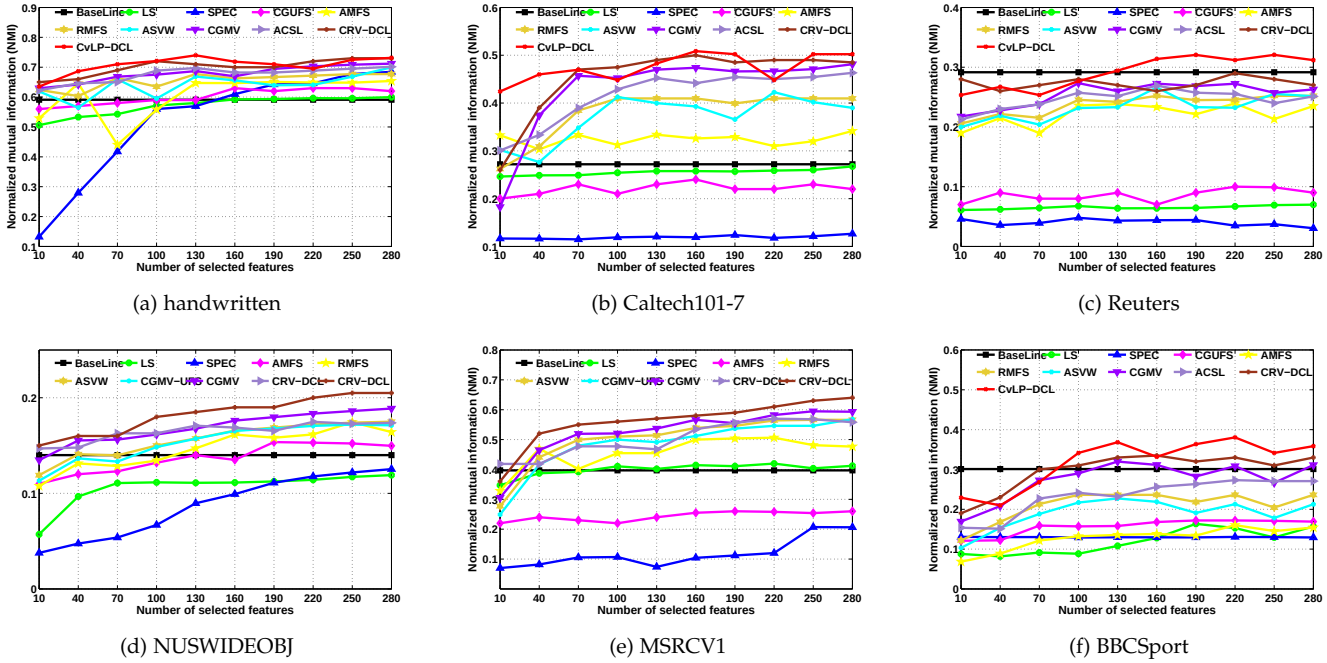
Fig. 1. The clustering accuracy (ACC) of using different selected features by different methods on different datasets.

proposed for human motion retrieval by using multiple features;

- **ASVW** [20]: Adaptive similarity learning with view weight for multi-view feature selection, which aims to learn a uniform similarity graph shared by different views to constrain the local structure of multi-view data;
- **RMFS** [21]: Robust multi-view feature selection,

TABLE 3
Clustering results (NMI% $\pm$ std%) of different algorithms on different datasets by implementing $K$-means and GMC on the selected features.

| Datasets | handwritten | | Caltech101-7 | | Reuters | | NUSWIDEOBJ | | MSRCV1 | | BBCSport | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | K-means | 59.11±1.89 | K-means | 27.19±1.00 | K-means | 29.16±2.51 | K-means | 14.00±0.17 | K-means | 39.69±2.40 | K-means | 30.10±1.28 |
| | GMC | 62.92 | GMC | 31.26 | GMC | 33.24 | GMC | 19.02 | GMC | 42.62 | GMC | 33.34 |
| LS | K-means | 59.97±1.44 | K-means | 26.36±1.07 | K-means | 7.63±0.91 | K-means | 12.13±0.18 | K-means | 42.63±4.01 | K-means | 16.79±6.54 |
| | GMC | 63.06 | GMC | 30.95 | GMC | 10.92 | GMC | 16.74 | GMC | 45.83 | GMC | 21.26 |
| SPEC | K-means | 68.45±3.98 | K-means | 12.35±1.06 | K-means | 6.04±0.00 | K-means | 12.82±0.19 | K-means | 22.30±5.14 | K-means | 13.24±0.06 |
| | GMC | 73.91 | GMC | 17.82 | GMC | 9.88 | GMC | 17.83 | GMC | 26.24 | GMC | 18.71 |
| CGUFS | K-means | 63.27±1.66 | K-means | 24.47±1.08 | K-means | 10.34±0.86 | K-means | 15.36±0.21 | K-means | 26.84±4.92 | K-means | 17.27±1.31 |
| | GMC | 66.97 | GMC | 29.55 | GMC | 16.58 | GMC | 18.36 | GMC | 34.19 | GMC | 20.11 |
| AMFS | K-means | 65.09±0.64 | K-means | 35.53±2.03 | K-means | 24.30±0.94 | K-means | 16.51±0.17 | K-means | 50.37±4.80 | K-means | 19.86±3.37 |
| | GMC | 70.52 | GMC | 38.94 | GMC | 29.68 | GMC | 20.62 | GMC | 54.27 | GMC | 22.77 |
| RMFS | K-means | 67.75±1.60 | K-means | 40.97±1.69 | K-means | 25.21±1.19 | K-means | 16.58±0.26 | K-means | 56.61±3.17 | K-means | 23.62±1.23 |
| | GMC | 72.43 | GMC | 43.27 | GMC | 29.63 | GMC | 22.77 | GMC | 60.06 | GMC | 27.87 |
| ASVW | K-means | 68.92±1.37 | K-means | 46.41±1.92 | K-means | 26.75±1.27 | K-means | 16.87±0.21 | K-means | 57.20±3.61 | K-means | 27.29±2.54 |
| | GMC | 74.52 | GMC | 50.76 | GMC | 30.37 | GMC | 22.28 | GMC | 61.73 | GMC | 32.58 |
| CGMV-UFS | K-means | 71.83±2.18 | K-means | 48.71±3.33 | K-means | 27.76±1.06 | K-means | 18.96±0.19 | K-means | 60.50±5.46 | K-means | 31.94±1.39 |
| | GMC | 75.67 | GMC | 52.37 | GMC | 32.51 | GMC | 23.54 | GMC | 65.31 | GMC | 35.49 |
| ACSL | K-means | 70.23±5.41 | K-means | 47.39±2.09 | K-means | 26.18±1.14 | K-means | 17.28±0.35 | K-means | 59.58±4.76 | K-means | 28.57±1.19 |
| | GMC | 74.29 | GMC | 51.94 | GMC | 31.36 | GMC | 23.02 | GMC | 63.69 | GMC | 33.42 |
| CRV-DCL | K-means | 72.65±2.20 | K-means | 49.86±3.14 | K-means | 29.14±1.02 | K-means | 19.76±0.23 | K-means | 62.36±5.38 | K-means | 32.41±1.35 |
| | GMC | 77.22 | GMC | 53.16 | GMC | 33.29 | GMC | 24.51 | GMC | 66.25 | GMC | 37.3 |
| CvLP-DCL | K-means | 74.11±2.03 | K-means | 51.71±2.49 | K-means | 32.22±1.73 | K-means | 21.84±0.41 | K-means | 64.07±5.23 | K-means | 34.62±1.74 |
| | GMC | 79.06 | GMC | 56.88 | GMC | 36.72 | GMC | 25.75 | GMC | 67.49 | GMC | 39.47 |




(a) handwritten (b) Caltech101-7 (c) Reuters

(d) NUSWIDEOBJ (e) MSRCV1 (f) BBCSport

Fig. 2. The normalized mutual information (NMI) of using different selected features by different methods on different datasets.

which applies robust multi-view k-means to generate pseudo labels, and the labels are used to regularize sparse feature selection;

- **CGMV-UFS** [22]: Consensus learning guided MV-UFS, which aims to learn the consensus cluster indicator matrix of multiple views by using the non-negative matrix factorization;
- **ACSL** [41]: Adaptive collaborative similarity learning for MV-UFS, which dynamically learns the collaborative similarity structure, and the similarity learning and feature selection are integrated into a unified framework.
- **CRV-DCL** [23]: The previous model proposed in our AAAI 2019 version.

There are several parameters need to be set in CvLP-DCL as well as other methods. For LS, SPEC, CGMV-UFS, CRV-DCL, ACSL and CvLP-DCL, the neighborhood size for constructing the intra-view similarity graph is set to 5 For AMFS, $r$ is set to 2 as suggested in the corresponding paper. For ASVW, we turn the regularization parameter $\lambda$ in $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. As to $\lambda$, $\alpha$ and $\beta$ in CvLP-DCL, we also tune their values by a "grid-search" strategy from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. For other parameters in different methods, they are set to default values or tuned as suggested in the original papers to obtain the optimal results. For each method, the best results by tuning the parameters are reported for fair comparison.

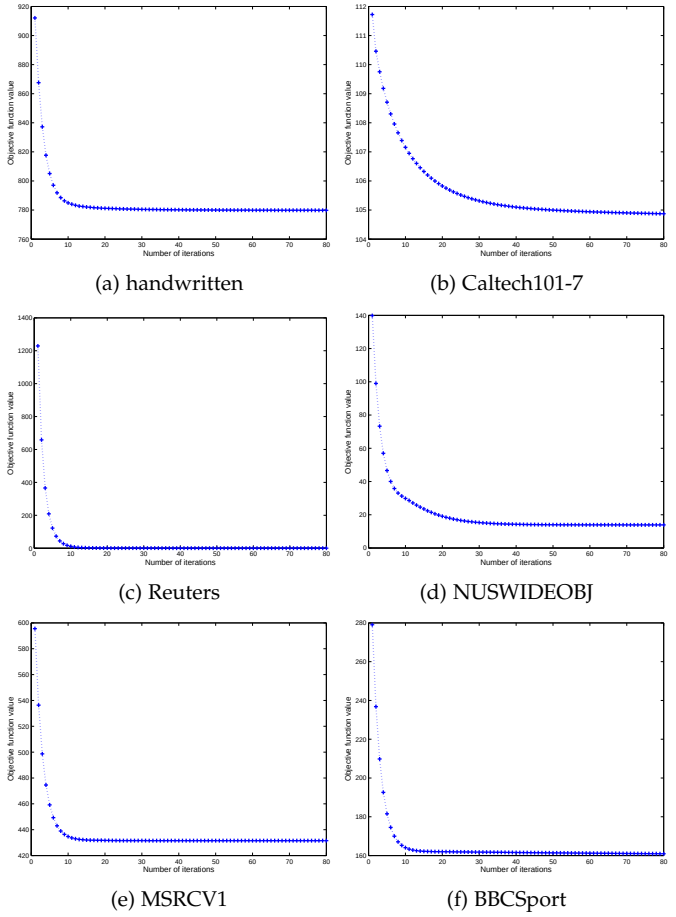
Since the optimal number of selected features for a certain dataset is hard to determine, we vary the number of selected features in a range for all datasets and report the best clustering results for each method. For all of the datasets, the selected feature numbers are varied from

$\{10, 40, 70, \cdots, 250, 280\}$. After obtaining the feature subsets, we implement two clustering methods on the selected features, i.e., the commonly used $K$-means and a recently proposed graph based multi-view clustering method, GMC [42]. Since $K$-means is sensitive to the initialization, we run it 20 times on the selected feature subsets with random starting points for eliminating the bias of initialization. Then, the average results with standard deviation of the 20 times running of $K$-means are recorded and reported. As to GMC, we use the recommended settings as reported in original paper for each dataset to obtain the final results.

### 4.3 Experimental Results

The ACC and NMI of different methods on different datasets are summarized in Table 2 and Table 3, respectively. From the results, we can see that the proposed CvLP-DCL consistently performs the best on all of the datasets when compared with other methods. As to handwritten, Caltech101-7 and MSRCV1, our method outperforms the baseline with more than 20% in terms of both ACC and NMI by using the $K$-means algorithm on the selected feature subsets. As to NUSWIDEOBJ, CvLP-DCL also obtains 5% improvement than the baseline. For Reuters and BBCSport, CvLP-DCL still outperforms all of other methods including the baseline. Therefore, the results validate the superiority of the proposed CvLP-DCL when compared with other methods. With a small subset of selected features, CvLP-DCL obtains better clustering results. In addition, compared with traditional single view unsupervised feature selection methods, the multi-view methods perform significantly better. We can see that CvLP-DCL can get more than 10% improvements in average when compared to the best result of all the other single-view methods. This is caused by the fact that single view methods characterize the structures of each data view independently and combine them by simply stacking. CvLP-DCL also consistently outperforms CRV-DCL on all datasets, which also demonstrates the efficacy of the cross-view similarity graph learning and regularization strategy.

As far as we know, there is no success way to determine the optimal number of selected features. Therefore, in order to illustrate the effect of feature selection to clustering, we show the clustering performance of different algorithms with respect to different numbers of selected features on different datasets. In Figure 1 and Figure 2, we plot the ACC and the NMI values with respect to the numbers of selected features on different datasets, respectively. As can be seen from the results, the proposed method can steadily perform better than other methods over a range of selected features. It should be noted that when using fewer features, our method can obtain higher clustering accuracy than the baseline excluding the Reuters and BBCSport datasets, which demonstrates that the selected subset of the features can not only reduce the computation cost, but also improve the clustering performance. As to Reuters and BBCSport, when the number of selected features is fewer than 80, our method dose not perform better than the baseline. However, when we select more features, the proposed CvLP-DCL can steadily perform better than other methods.




(a) handwritten (b) Caltech101-7

(c) Reuters (d) NUSWIDEOBJ

(e) MSRCV1 (f) BBCSport

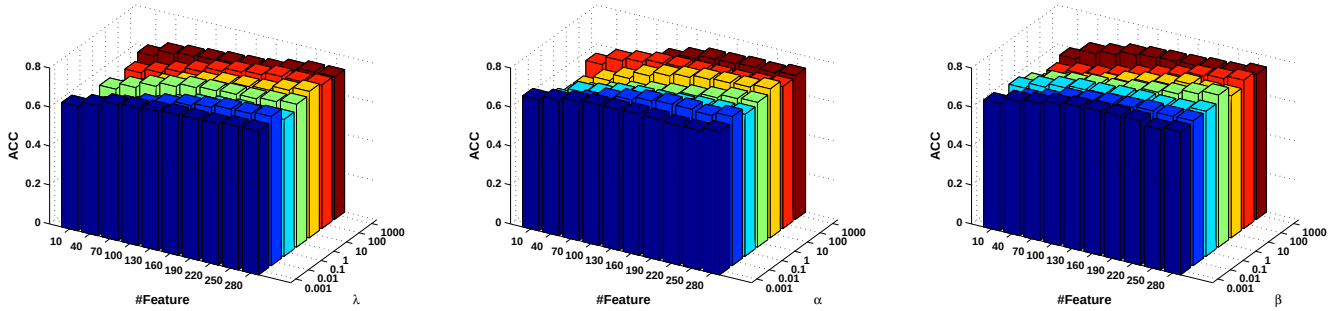
Fig. 4. Objective function values of Eq. (7) with varying iteration times on different datasets.

### 4.4 Parameter Sensitivity

There are three parameters in our model as formulated by Eq. (7) (i.e., $\lambda$, $\alpha$ and $\beta$). To further demonstrate the performance of the proposed method, we study its sensitivity w.r.t. different parameters. Due to the page space limitation, we only report the ACC and NMI of handwritten dataset by using the $k$-means algorithm. At each time, we fix two parameters and show the performance of our method by varying the rest one parameter. Figure 3 plots the ACC and NMI values given by CvLP-DCL for different $\lambda$, $\alpha$, $\beta$ and selected features. The experimental results show that our CvLP-DCL is not very sensitive to the three hyper-parameters, but it is relatively sensitive to the number of selected features. However, this is a common problem for most unsupervised feature selection methods since the optimal feature set for each dataset is hard to determine.

### 4.5 Efficacy of the Learned Uniform Similarity Graph

In our proposed model as described by Eq. (7), we learn a cross-view similarity graph $\bar{\mathbf{S}}$ from multiple view-specific similarity graphs $\mathbf{S}^v (v = 1, \ldots, V)$, and then $\bar{\mathbf{S}}$ is used to regularize the local geometrical structure of original data in the label space. In order to validate the efficacy of the cross-view similarity graph learning process, we take the handwritten dataset as an example and intuitively show the initial similarity graphs of the 2nd, 4th and 6th view ($\mathbf{S}^2$, $\mathbf{S}^4$ and $\mathbf{S}^6$), and the learned cross-view similarity graph ($\bar{\mathbf{S}}$)




(a) $\alpha = \beta = 1$ (b) $\lambda = \beta = 1$ (c) $\lambda = \alpha = 1$

(d) $\alpha = \beta = 1$ (e) $\lambda = \beta = 1$ (f) $\lambda = \alpha = 1$

Fig. 3. ACC and NMI of CvLP-DCL with different $\lambda$, $\alpha$, $\beta$, and number of selected features on handwritten dataset. The first row presents the ACC with varying parameters and the second row shows the NMI with varying parameters.

in Figure 5. As can be seen, there are some noisy values in the initial view-specific similarity graphs. After the learning process, the noisy values can be effectively removed and a cleaner cross-view similarity graph with diagonal structure is obtained. As a result, the learned cross-view similarity graph can better preserve the locality of original data.

### 4.6 Empirically Convergence Validation

In Section 3.5, we theoretically analyse the convergence property of Algorithm 1. In this section, we empirically validate its convergence property. As shown in Figure 4, we plot the objective function values of Eq. (7) with varying iteration times on different datasets ($\lambda$, $\alpha$ and $\beta$ are fixed to 1), the results show that Algorithm 1 converges very fast and the objective function value goes stable almost within 20 iterations.

## 5 Conclusions

This paper introduces a novel multi-view unsupervised feature selection method via cross-view local structure preserved diversity and consensus representation learning. The proposed method captures both the common information and distinguishing knowledge across different views by projecting each view of original data into a common label space, which is composed of a consensus part and a diversity part. Meanwhile, in order to preserve the local structure of samples in the label space, multiple pre-defined view-specific similarity graphs are used to learn a shared similarity graph across different views. Experiments results with parameter sensitivity analysis on real-world multi-view datasets demonstrate the efficacy of the proposed method.

## Acknowledgment

This work was supported in part by the National Science Foundation of China under Grant 62076228, 61701451 and 41925007, and in part by Natural Natural Science Foundation of Hubei Province under Grant 2020CFB644, and in part by CAAI-Huawei MindSpore Open Fund.

## References

[1] D. G. Lowe and D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

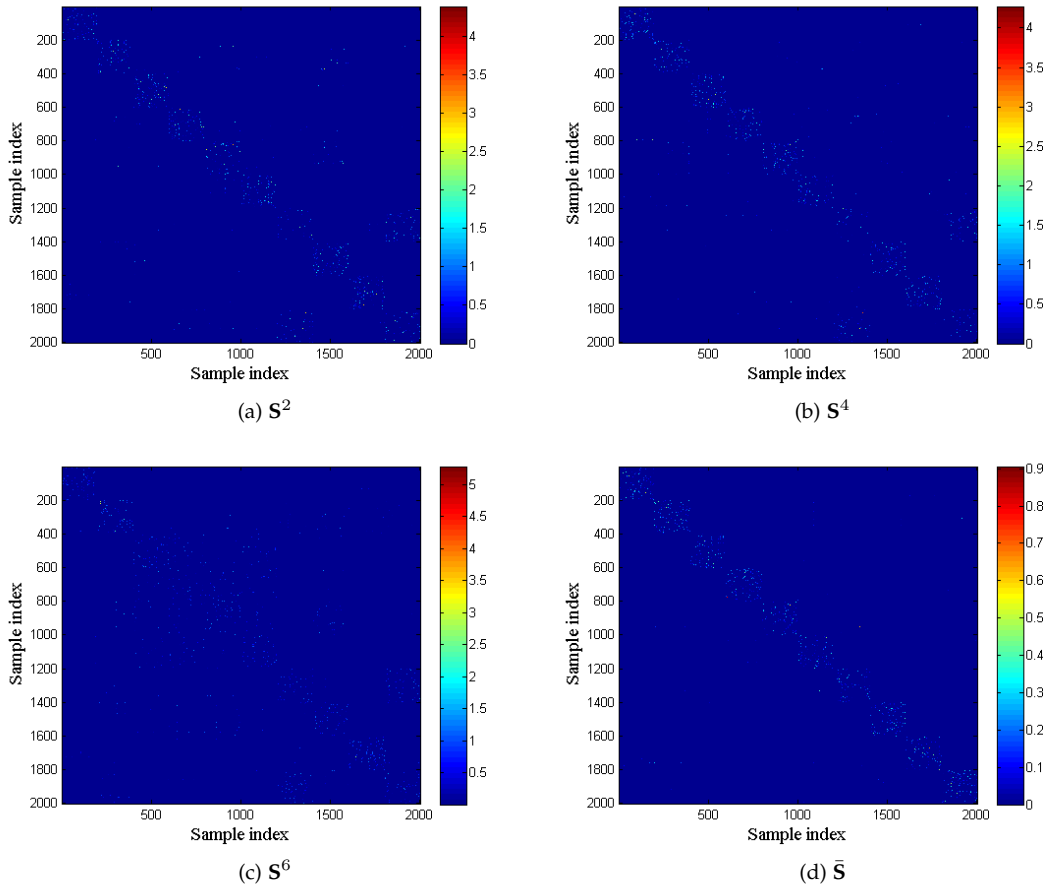
[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[4] L. Li, "Mpgraph: multi-view penalised graph clustering for predicting drug-target interactions," *Iet Systems Biology*, vol. 8, no. 2, pp. 67–73, 2014.

[5] L. Li and M. Cai, "Drug target prediction by multi-view low rank embedding." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. PP, no. 99, pp. 1–1, 2017.

[6] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao, "Latent multi-view subspace clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4333–4341.

[7] H. Zhao, Z. Ding, and Y. Fu, "Multi-view clustering via deep matrix factorization," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 2921–2927.




(a) $\mathbf{S}^2$ (b) $\mathbf{S}^4$

(c) $\mathbf{S}^6$ (d) $\bar{\mathbf{S}}$

Fig. 5. The view-specific similarity graphs and the learned cross-view similarity graph of different feature views of handwritten dataset (Zoom in for better visualization).

[8] W. Zhuge, F. Nie, C. Hou, and D. Yi, "Unsupervised single and multiple views feature extraction with structured graph," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2347–2359, 2017.

[9] H. Tao, C. Hou, D. Yi, and J. Zhu, "Multiview classification with cohesion and diversity," *IEEE transactions on cybernetics*, vol. 50, no. 5, pp. 2124 – 2137, 2018.

[10] Y. Wang, L. Wu, X. Lin, and J. Gao, "Multiview spectral clustering via structured low-rank matrix factorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, 2018.

[11] J. Wu, S. Pan, X. Zhu, C. Zhang, and P. S. Yu, "Multiple structure-view learning for graph classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 7, pp. 3236–3251, 2018.

[12] W. Zhuge, C. Hou, S. Peng, and D. Yi, "Joint consensus and diversity for multi-view semi-supervised classification," *Machine Learning*, pp. 1–21, 2019.

[13] J. H. Friedman, "On bias, variance, 0/1loss, and the curse-of-dimensionality," *Data Mining & Knowledge Discovery*, vol. 1, no. 1, pp. 55–77, 1997.

[14] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2005, pp. 507–514.

[15] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection," in *National Conference on Artificial Intelligence*, 2008, pp. 671–676.

[16] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *International Conference on Machine Learning*, 2007, pp. 1151–1157.

[17] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *AAAI Conference on Artificial Intelligence*, 2010.

[18] Z. Wang, Y. Feng, T. Qi, X. Yang, and J. J. Zhang, "Adaptive multi-view feature selection for human motion retrieval," *Signal Processing*, vol. 120, pp. 691–701, 2016.

[19] Y. Feng, J. Xiao, Y. Zhuang, and X. Liu, "Adaptive unsupervised multi-view feature selection for visual concept recognition," in *Asian Conference on Computer Vision*, 2012, pp. 343–357.

[20] C. Hou, F. Nie, H. Tao, and D. Yi, "Multi-view unsupervised feature selection with adaptive similarity and view weight," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1998 – 2011, 2017.

[21] H. Liu, H. Mao, and Y. Fu, "Robust multi-view feature selection," in *IEEE International Conference on Data Mining*, 2017, pp. 281–290.

[22] C. Tang, J. Chen, X. Liu, M. Li, P. Wang, M. Wang, and P. Lu, "Consensus learning guided multi-view unsupervised feature selection," *Knowledge-Based Systems*, vol. 160, pp. 49–60, 2018.

[23] C. Tang, X. Zhu, X. Liu, and L. Wang, "Cross-view local structure preserved diversity and consensus learning for multi-view unsupervised feature selection," in *AAAI Conference on Artificial Intelligence*, 2019, pp. 5101–5108.

[24] Y. Wan, S. Sun, and C. Zeng, "Adaptive similarity embedding for unsupervised multi-view feature selection," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[25] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu, "Global and local structure preservation for feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 6, pp. 1083–1095, 2014.

[26] F. Nie, Z. Wei, and X. Li, "Unsupervised feature selection with structured graph optimization," in *AAAI Conference on Artificial Intelligence*, 2016, pp. 1302–1308.

[27] C. Tang, L. Cao, X. Zheng, and M. Wang, "Gene selection for microarray data classification via subspace learning and manifold regularization." *Medical & Biological Engineering & Computing*, no. 6871, pp. 1–14, 2017.

[28] C. Tang, X. Zhu, J. Chen, P. Wang, X. Liu, and J. Tian, "Robust graph regularized unsupervised feature selection," *Expert Systems With Applications*, vol. 96, pp. 64–76, 2018.

[29] C. Tang, X. Liu, M. Li, P. Wang, J. Chen, L. Wang, and W. Li, "Robust unsupervised feature selection via dual self-representation

and manifold regularization," *Knowledge-Based Systems*, vol. 145, pp. 109–120, 2018.
[30] H. Liu, M. Shao, and Y. Fu, "Feature selection with unsupervised consensus guidance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2319–2331, 2018.
[31] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[32] K. Bache and M. Lichman, "Uci machine learning repository," 2013.
[33] F. F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *IEEE International Cpnference on Computer Vision and Pattern Recognition Workshop*, 2005, pp. 178–178.
[34] D. Dueck and B. J. Frey, "Non-metric affinity propagation for unsupervised image categorization," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.
[35] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *AAAI Conference on Artificial Intelligence*, 2015, pp. 2750–2756.
[36] M. R. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views – an application to multilingual text categorization," in *Advances in Neural Information Processing Systems*, 2009, pp. 28–36.
[37] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *ACM International Conference on Image and Video Retrieval*, 2009, p. 48.
[38] J. Xu, J. Han, and F. Nie, "Discriminatively embedded k-means for multi-view clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5356–5364.
[39] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition." in *AAAI Conference on Artificial Intelligence*, 2014, pp. 2149–2155.
[40] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 793–804, 2013.
[41] X. Dong, L. Zhu, X. Song, J. Li, and Z. Cheng, "Adaptive collaborative similarity learning for unsupervised multi-view feature selection," in *International Joint Conference on Artificial Intelligence*, 2018, pp. 2064–2070.
[42] H. Wang, Y. Yang, and B. Liu, "Gmc: Graph-based multi-view clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 6, pp. 1116–1129, 2019.

**Chang Tang (Member, IEEE)** received his Ph.D. degree from Tianjin University, Tianjin, China in 2016. He joined the AMRL Lab of the University of Wollongong between Sep. 2014 and Sep. 2015. He is now an associate professor at the School of Computer Science, China University of Geosciences, Wuhan, China. Dr. Tang has published 50+ peer-reviewed papers, including those in highly regarded journals and conferences such as IEEE T-PAMI, IEEE T-KDE, IEEE T-MM, IEEE T-HMS, IEEE SPL, AAAI, IJCAI, ICCV, CVPR, ACMM, ICME etc. He regularly served as the Technical Program Committees of top conferences such as NIPS, ICML, IJCAI, ICME, AAAI, ICCV, CVPR, etc. His current research interests include machine learning and computer vision.

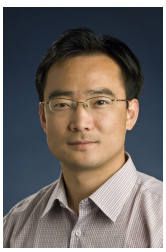
**Xiao Zheng** received her master degree from the Tianjin Medical University, Tianjin, China. She is currently pursuing the Ph.D. degree with the National University of Defense Technology, China. Her recent research interests include machine learning and medical data processing.

**Xinwang Liu (Senior Member, IEEE)** received his PhD degree from National University of Defense Technology (NUDT), China. He is now full professor of School of Computer, NUDT. His current research interests include kernel learning and unsupervised feature learning. Dr. Liu has published 60+ peer-reviewed papers, including those in highly regarded journals and conferences such as IEEE T-PAMI, IEEE T-KDE, IEEE T-IP, IEEE T-NNLS, IEEE T-MM, IEEE T-IFS, NeurIPS, ICCV, CVPR, AAAI, IJCAI, etc. He is a senior member of IEEE. More information can be found at xinwangliu.github.io.

**Wei Zhang (Member, IEEE)** received the B.E. degree from Zhejiang University in 2004, the M.S. degree from Liaoning University in 2008, and the Ph.D. degree from Shandong University of Science and Technology in 2018. He is currently an Associate Professor with the Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences). His research interests include future generation network architectures, edge computing and edge intelligence.

**Jing Zhang** received her Ph.D. degree at the Advanced Multimedia Research Laboratory from the University of Wollongong, Australia. She is now a lecturer in College of Software, Beihang University, Beijing, China. Her recent research interests include computer vision and machine learning.

**Jian Xiong (Member, IEEE)** (M'13) received the BS degree in engineering, and the MS and PhD degrees in management from the National University of Defense Technology, Changsha, China, in 2005, 2007, and 2012, respectively. He is an associate professor with the School of Business Administration, Southwestern University of Finance and Economics. His research interests include data mining, multi objective evolutionary optimization, multiobjective decision making, project planning, and scheduling. He is a member of the IEEE.



**Lizhe Wang (Fellow, IEEE)** received the B.E. and M.E. degrees from Tsinghua University, Beijing, China, and the Dr.Eng. (Magna Cum Laude) degree from the University of Karlsruhe, Karlsruhe, Germany.

He is currently a ChuTian Chair Professor with the School of Computer Science, China University of Geosciences, Beijing, and also a Professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing. His research interests include high-performance computing, eScience, and remote sensing image processing. Dr. Wang is a fellow of the IET and the British Computer Society. He serves as an Associate Editor of the IEEE Transactions on Parallel and Distributed Systems, the IEEE Transactions on Cloud Computing, and the IEEE Transactions on Sustainable Computing.