Neurocomputing 656 (2025) 131487

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Multi-view unsupervised feature selection based on graph discrepancy learning

Yiwan Xu [a, b], Xijiong Xie [a, b,*], Xianliang Jiang [a, b], Yujie Xiong [c]

[a] School of Information Science and Engineering, Ningbo University, Ningbo, 315211, China
[b] Key Laboratory of Mobile Network Application Technology of Zhejiang Province, Ningbo University, 315211, Ningbo, China
[c] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, 201620, China

ARTICLE INFO

Communicated by G. Yu

Keywords:
Local and global structure
Graph discrepancy
Kernel mapping
Multi-view learning
Unsupervised feature selection

ABSTRACT

In multi-view learning, unsupervised feature selection plays a vital role in reducing dimensionality while preserving discriminative information distributed across diverse data modalities. Despite notable progress, existing approaches frequently exhibit two key limitations: they often overlook the complementary benefits of integrating global and local structural information, and they inadequately model complex nonlinear relationships or align structural representations across views. To address these challenges, we propose a novel framework, termed Multi-view unsupervised feature selection based on graph discrepancy learning (GDFS). The proposed method jointly constructs global graph structures in a projected low-dimensional space and local graphs in a nonlinear kernel-induced space, effectively capturing both high-level semantic structures and fine-grained neighborhood dependencies. A graph discrepancy term is introduced to explicitly reduce structural discrepancies between global and local representations, thus enhancing consistency and robustness. In addition, a low-rank tensor constraint is applied to the stack of global graphs to uncover high-order correlations across views. A consensus clustering matrix is further learned to provide pseudo-label supervision, which guides the selection of discriminative features. Extensive experiments on six benchmark multi-view datasets demonstrate that GDFS consistently surpasses state-of-the-art methods in terms of clustering performance, thereby confirming its effectiveness, scalability, and generalizability. The code is available at https://github.com/xyw0111/2025-GDFS.

1. Introduction

Multi-view data encapsulates multiple heterogeneous yet complementary perspectives of the same underlying entity, obtained through diverse sensors, feature extraction algorithms, or different observational angles. This data paradigm has become increasingly prevalent across a wide range of domains. For instance, in image analysis, features such as Scale-Invariant Feature Transform (SIFT) [1], Histogram of Oriented Gradients (HOG) [2], and Local Binary Patterns (LBP) [3] offer distinct characterizations of visual content, each capturing different aspects of image structure. In text mining, multilingual documents provide semantically varied representations, and in human activity recognition [4], modalities such as RGB imagery, depth sensing, and wearable devices contribute diverse streams of behavioral information. Compared with single-view datasets, multi-view data often contain richer and complementary information, thereby improving the effectiveness of downstream learning tasks.

Nonetheless, the high dimensionality typically associated with multi-view data poses considerable computational and storage challenges. From a data perspective [5], the fundamental challenge lies in balancing complementarity and consistency across views while preserving key information. To address these issues, dimensionality reduction techniques are commonly employed, which can be categorized into two main approaches: feature extraction [6–8], which maps original features into new lower-dimensional representations, and feature selection [9,10], which directly identifies salient features by removing redundancies. Among these, feature selection has attracted growing attention [11,12] due to its unique advantages: (1) maintaining the semantic integrity of original features [13]; (2) offering better interpretability. Recent advances in consensus learning [14] and complementary frameworks [15] have further enhanced feature selection's ability to handle view consistency while preserving interpretability.

* Corresponding author at: School of Information Science and Engineering, Ningbo University, Ningbo, 315211, China.
Email addresses: 2311100094@nbu.edu.cn (Y. Xu), xjxie11@gmail.com (X. Xie), jiangxianliang@nbu.edu.cn (X. Jiang), xiong@sues.edu.com (Y. Xiong).

https://doi.org/10.1016/j.neucom.2025.131487
Received 4 June 2025; Received in revised form 8 August 2025; Accepted 6 September 2025
Available online 8 September 2025

Within this context, multi-view feature selection has emerged as a vital approach, which offers both interpretability and robustness in handling the complexity of multi-view data. Depending on the availability of supervision, existing methods are typically classified into supervised [16,17], semi-supervised [18–20], and unsupervised [21–27] strategies. Supervised methods leverage annotated labels to guide the selection of discriminative features, while semi-supervised methods benefit from both labeled and unlabeled data. In contrast, unsupervised methods operate without any form of supervision, relying instead on intrinsic structural properties of the data—a particularly challenging scenario, yet highly applicable given the scarcity of labeled data in many real-world applications.

Despite notable progress, existing unsupervised multi-view feature selection techniques face several critical limitations. Many approaches emphasize either the preservation of global structure [28] or the modeling of local neighborhood relationships [29], and often neglect the synergistic interplay between these two structural perspectives. This dichotomy can lead to incomplete exploitation of the rich structural information inherent in multi-view data. Furthermore, a substantial number of methods rely on linear assumptions, which are insufficient for modeling the complex nonlinear relationships commonly encountered in practical scenarios. To address this, kernel mapping [30] has emerged as an effective strategy, as it projects these samples into a high-dimensional Reproducing Kernel Hilbert Space, thereby enhancing the model's ability to capture nonlinear relationships. An additional yet often overlooked challenge lies in achieving structural alignment across different views. Without explicit modeling of inter-view consistency, selected features may lack coherence, diminishing their utility. To this end, we introduce a graph discrepancy term to quantify and minimize the inconsistency between local and global structures, thereby promoting structural coherence.

In this work, we introduce a novel framework, termed Multi-view unsupervised feature selection based on graph discrepancy learning (GDFS). The overall architecture is depicted in Fig. 1. For each view, local graph structures are constructed in a nonlinear kernel-induced space to effectively capture complex neighborhood relationships. In parallel, global graph structures are learned from low-dimensional projections, which enhance both feature discriminability and structural integrity. A graph discrepancy term is then introduced to jointly optimize local and global representations. To uncover shared higher-order relationships across views, all global graphs are stacked into a third-order tensor, and a tensor nuclear norm is applied to extract common latent correlations. Furthermore, a consensus clustering matrix is learned to maintain consistency across views and to provide discriminative pseudo-supervision for feature selection. Collectively, these components constitute a comprehensive and cohesive framework that advances unsupervised feature selection by integrating local-global graph modeling, structural alignment, and multi-view learning.

The core contributions of this study are summarized as follows.

1. We present an integrated framework that concurrently constructs local and global graph structures to capture the intricate neighborhood relationships and overall distribution, respectively. These dual representations are coupled via a graph discrepancy term, which effectively enhances structural coherence and model robustness.
2. To better accommodate the nonlinear nature of real-world data, local graphs are constructed within a kernel-induced space. This design allows the model to effectively capture complex nonlinear relationships, particularly those characterized by curved decision boundaries, which conventional linear graph constructions are often unable to represent accurately.
3. An efficient optimization algorithm is developed to solve the proposed objective function, which offers both stable convergence and manageable computational complexity. The efficacy and practical value of the proposed approach are validated through extensive clustering experiments performed on a broad spectrum of benchmark multi-view datasets.

The structure of this paper is outlined as follows. Section 2 provides a concise review of methods related to multi-view unsupervised feature selection. Section 3 introduces our multi-view unsupervised feature selection approach, and also discusses its optimization, convergence, and

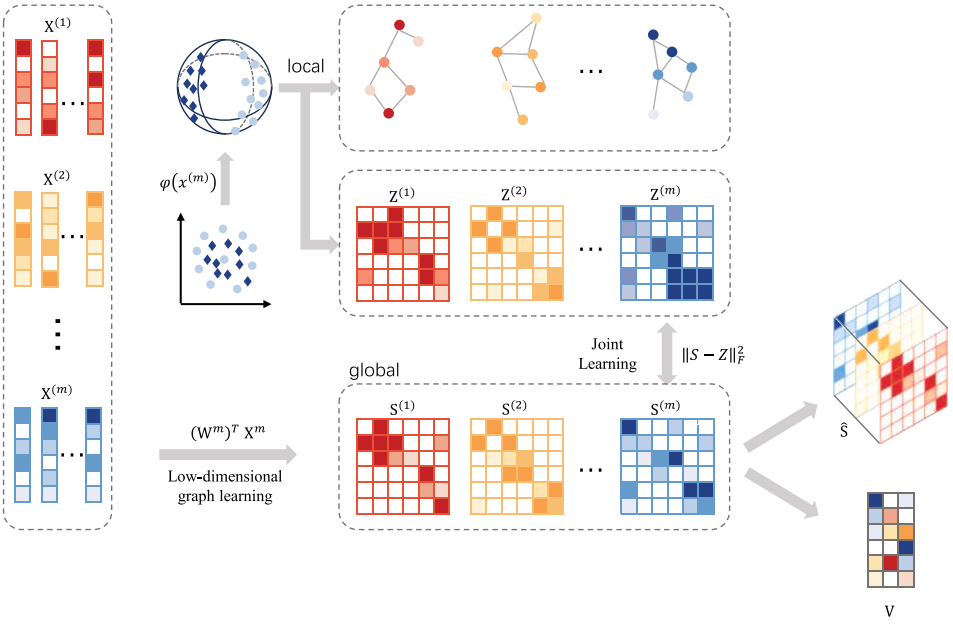


Fig. 1. The overall framework of the proposed method. On the left, multi-view raw features are used to construct local graphs via kernel mapping (top) and global graphs via low-dimensional projection (bottom). The local graphs $\mathbf{Z}^{(m)}$ and global graphs $\mathbf{S}^{(m)}$ are jointly aligned through a graph discrepancy term to enhance structural consistency. All graphs are stacked into a third-order tensor with low-rank regularization to extract cross-view correlations. A consensus clustering matrix $\mathbf{V}$ is learned to further unify graph structures and guide unsupervised feature selection.

computational complexity. Section 4 presents experimental results on multiple multi-view datasets. Finally, Section 5 concludes the study.

2. Related work

2.1. Kernel generation

Given a set of data samples $\{\mathbf{x}_i\}_{i=1}^N$ drawn from an input space $\mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$, kernel methods project these samples into a high-dimensional Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H} \subseteq \mathbb{R}^{d_\mathcal{H}}$ via an implicit feature mapping $\varphi(\cdot)$. Due to the potentially infinite dimensionality of $\mathcal{H}$, this mapping is typically not explicitly defined, and renders direct computation of embedded representations intractable. Fortunately, Mercer's theorem [31] enables the computation of inner products in $\mathcal{H}$ through a kernel function $k(\cdot,\cdot)$ operating in the original input space $\mathcal{X}$, formulated as:

$$\mathbf{K}[i,j] = \varphi(\mathbf{x}_i)^\top \varphi(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j), \tag{1}$$

where $\mathbf{K}[i,j]$ represents the (i,j)-th entry of the kernel matrix $\mathbf{K}$. A list of commonly used kernel functions is provided in Table 1.

In multi-view learning scenarios, current multi-kernel learning techniques often construct one or more kernels per view, with the primary objective of improving performance by designing more effective fusion strategies to aggregate discriminative information across views [32–36]. However, these approaches frequently underestimate a critical limitation—namely, that the quality of individual kernel often poses a bottleneck to overall performance. Although several studies have investigated optimal parameter selection in kernel functions [37], this topic falls outside the scope of our study. Instead, we introduce a novel kernel generation paradigm, referred to as Cross-view Multiple Kernels (CMK), which retains the form and parameterization of traditional kernel functions while offering a principled and structurally coherent approach to kernel construction, specifically tailored to multi-view settings.

2.2. Unsupervised feature selection via low-rank tensor-based graph learning

In the absence of supervisory information, the inherent distributional patterns within data offer a valuable foundation for guiding unsupervised feature selection. Consequently, a wide array of methodologies has been developed to identify feature subsets that most effectively unveil the underlying structural characteristics of the data [38]. Conventional approaches typically employ a two-step procedure: first estimating the intrinsic structure using the complete set of input features, and subsequently selecting those features that best preserve this structure. However, such strategies are prone to degradation in performance when the estimated structure is compromised by noise, redundancy, or irrelevant variables.

To address these challenges, recent research has proposed integrating graph learning and feature selection within a unified optimization framework, thereby enabling mutual enhancement between the two processes throughout the learning procedure [39]. Building on this foundation, multi-view learning scenarios have inspired the introduction of low-rank tensor regularization to better capture high-order relationships across heterogeneous feature spaces. Specifically, by assembling the view-specific similarity graphs or selection matrices into a high-order tensor and imposing a low-rank constraint, these methods are capable of capturing both shared and complementary information across views [40]. This approach can be mathematically formulated as follows:

$$\begin{aligned} &\min_{\mathbf{S},\mathbf{W}_v} \sum_{i=1}^{n}\sum_{j=1}^{n} \left\| \mathbf{W}_v^\top \mathbf{X}_{v[:,i]} - \mathbf{W}_v^\top \mathbf{X}_{v[:,j]} \right\|^2 S_{ij}^{(v)} + \Omega(\mathbf{W}) + \|\mathbf{S}\|_\otimes \\ &\text{s.t.} \mathbf{W}_v^\top \mathbf{W}_v = \mathbf{I}, \sum_{j=1}^{n} S_{ij}^{(v)} = 1, S_{ij}^{(v)} \geq 0, \\ &\quad \mathbf{S} = \Phi(\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \ldots, \mathbf{S}^{(V)}), S_{ii}^{(v)} = 0. \end{aligned} \tag{2}$$

While such models have shown promising results, they are not without limitations. First, their reliance on linear projection matrices $\mathbf{W}_v$ restricts their ability to model complex and nonlinear feature relationships—an essential characteristic of many real-world datasets. Second, these methods predominantly emphasize the preservation of local structural information, and often neglect global data dependencies which are equally critical for comprehensive and robust feature selection.

3. Methodology

3.1. Preliminary

In this paper, matrices, vectors, and scalars are represented by bold uppercase letters, bold lowercase letters, and normal italic letters, respectively. n, v, and k denote the number of samples, the number of views, and the number of clusters, respectively. $\mathcal{I}_k$ represents a $k \times k$ identity matrix. The data matrix of the mth view is defined as $\mathbf{X}^m \in \mathbb{R}^{d_m \times n}$. $\mathrm{Tr}(\cdot)$ and $\mathrm{diag}(\cdot)$ represent the trace operator and diagonal elements of a matrix, respectively. The Frobenius norm and $l_{2,1}$-norm are denoted as $\|\cdot\|_F$ and $\|\cdot\|_{2,1}$, respectively. Additionally, third-order tensors are represented by bold calligraphic letters, such as $\mathcal{J} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. The t-SVD-based tensor nuclear norm of $\mathcal{J}$ is defined as $\|\mathcal{J}\|_\otimes = \sum_{m=1}^{n_3} \|\mathcal{J}_f^{(m)}\|_* = \sum_{i=1}^{\min(n_1,n_2)} \sum_{m=1}^{n_3} |S_f^{(m)}(i,i)|$, where $S_f^{(m)}$ is obtained by the SVD on all front slices of $\mathcal{J}_f$. Table 2 summarizes the main notations used in this article.

Definition 1. The $l_{2,1}$-norm [41] is defined as

$$\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{d} \|\mathbf{w}_{i,:}\|_2, \tag{3}$$

which computes the sum of the ℓ_2-norms of all rows in $\mathbf{W}$. This formulation encourages row-wise sparsity, i.e., entire rows of $\mathbf{W}$ to become zero.

Table 1
Representatives of traditional kernel functions.

Kernel type	Formulation
Linear	$\alpha \mathbf{x}_i^\top \mathbf{x}_j + c$
Gaussian	$\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2)$
Polynomial	$(\alpha \mathbf{x}_i^\top \mathbf{x}_j + c)^d$
Sigmoid	$\tanh(\alpha \mathbf{x}_i^\top \mathbf{x}_j + c)$
Cauchy	$(\|\mathbf{x}_i - \mathbf{x}_j\|^2 / \sigma + 1)^{-1}$

Table 2
Summary of notations used in this article.

Notations	Descriptions
$\mathbf{X}^m \in \mathbb{R}^{d_m \times n}$	Data matrix of the mth view
$\mathbf{W}^m \in \mathbb{R}^{d_m \times k}$	Feature weight matrix of the mth view
$\mathbf{S}^m \in \mathbb{R}^{n \times n}$	Local similarity matrix of the mth view
$\mathbf{Z}^m \in \mathbb{R}^{n \times n}$	Global similarity matrix of the mth view
$\mathbf{V} \in \mathbb{R}^{n \times k}$	The consensus clustering result
$\mathcal{J} \in \mathbb{R}^{n \times n \times v}$	An auxiliary tensor variable
$\mathcal{M} \in \mathbb{R}^{n \times n \times v}$	Lagrange multiplier
$\mathbf{I}_k$	An $k \times k$ identity matrix
d_m	The number of features in the mth view
u	Penalty factor
n	The number of instances
k	The number of clusters
v	The number of views
$\alpha, \beta, \gamma, \rho$	Balancing parameters

3.2. The proposed GDFS model

3.2.1. Joint global and local graph learning

In unsupervised learning tasks such as clustering and representation learning, a central challenge is the accurate recovery of the global structure embedded within high-dimensional data. Traditional methods typically rely on the full feature space to infer inter-sample affinities; however, the presence of noisy, irrelevant, or redundant features often degrades the quality of the constructed similarity graph, and ultimately impairs downstream performance.

To address this limitation, recent advances advocate a unified optimization framework that simultaneously performs global graph construction and feature selection. This joint modeling paradigm enables mutual enhancement between the two components throughout the training process. A representative formulation is given by

$$\min_{\mathbf{W},\mathbf{S}} \left\|\mathbf{W}^T\mathbf{X} - \mathbf{W}^T\mathbf{X}\mathbf{S}\right\|_F^2 + \Omega(\mathbf{W}), \quad \text{s.t.} \quad \mathbf{S} \in C. \tag{4}$$

where $\mathbf{S} \in \mathbb{R}^{n\times n}$ denotes a global similarity matrix capturing reconstruction-based relationships among samples, and $\mathbf{W} \in \mathbb{R}^{d\times k}$ serves as a feature weighting matrix that facilitates the selection of discriminative features while preserving the global data structure. The constraint set C and the regularizer $\Omega(\mathbf{W})$ are designed to promote desired structural properties, such as sparsity or orthogonality. This formulation extends naturally to multi-view learning framework.

$$\begin{aligned}&\min_{\mathbf{W}^m,\mathbf{S}^m} \sum_{m=1}^{v} \left(\|(\mathbf{W}^m)^T\mathbf{X}^m - (\mathbf{W}^m)^T\mathbf{X}^m\mathbf{S}^m\|_F^2 + \gamma\|\mathbf{W}^m\|_{2,1} \right) \\ &\text{s.t. } (\mathbf{W}^m)^T\mathbf{X}^m(\mathbf{X}^m)^T\mathbf{W}^m = I_k.\end{aligned} \tag{5}$$

From a local perspective, preserving fine-grained neighborhood structures is equally essential. Classical graph construction methods, such as k-nearest neighbors, often rely on fixed hyperparameters and Euclidean distances, which make them unsuitable for complex data geometries and sensitive to noise. To address these issues, adaptive neighborhood learning has been introduced.

$$\min_{\mathbf{z}_i} \sum_{j=1}^{n} \|\mathbf{x}_i - \mathbf{x}_j\|^2 z_{ij} \quad \text{s.t.} \quad \mathbf{z}_i^T\mathbf{1} = 1, \tag{6}$$

where z_{ij} denotes the similarity between sample i and sample j. The term $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ measures the Euclidean distance between them. A smaller distance leads to a larger value of z_{ij}, which encourages the model to assign higher similarity scores to truly relevant neighbors. By generalizing single-view subspace learning to multi-view cases, Eq. (6) is further expressed in the following matrix form.

$$\min_{\mathbf{z}_i^{(m)}} \sum_{m=1}^{v}\sum_{i,j=1}^{n} \|\mathbf{x}_i^m - \mathbf{x}_j^m\|^2 z_{ij}^m \quad \text{s.t.} \quad (\mathbf{z}_i^m)^T\mathbf{1} = 1. \tag{7}$$

Moreover, local graph learning focuses on preserving fine-grained geometric relationships among neighboring data points. Real-world datasets commonly exhibit complex nonlinear patterns at the local level, which are difficult to model accurately in the original space. Kernel mapping helps to linearize these nonlinear local structures by projecting the data into a high-dimensional space, where local neighborhood relationships become more distinguishable. As such, applying kernel mapping in local graph learning facilitates the construction of more expressive and discriminative neighborhood graphs, thereby enhancing the effectiveness of feature selection. Kernel extension of Eq. (7) can be obtained as

$$\begin{aligned}&\min_{\mathbf{Z}^m} \rho \sum_{m=1}^{v}\sum_{i,j=1}^{n} \|\phi^m(X_i^m) - \phi^m(X_j^m)\|_2^2 Z_{ij}^m \\ &\text{s.t.} diag(\mathbf{Z}^m) = 0, (\mathbf{z}_i^m)^T\mathbf{1} = 1,\end{aligned} \tag{8}$$

where the function $\phi^m(\cdot)$ represents nonlinear mapping. Conventional graph learning often focuses on either global or local structures, which limits its ability to model complex multi-view data. We propose a unified framework that integrates global graph learning with kernel-based adaptive neighborhood modeling to capture multi-level structures.

Specifically, global graph learning focuses on modeling the overall data distribution and typically constructs the similarity graph directly in the original feature space. This approach is computationally efficient and robust, which makes kernel mapping generally unnecessary. In contrast, local graph learning emphasizes fine-grained neighborhood structures. By projecting data into a high-dimensional space through kernel mapping, it becomes possible to uncover latent nonlinear relationships, thereby enhancing the expressiveness and discriminative power of the learned adjacency graph. Our method can be formulated as

$$\begin{aligned}&\min_{\mathbf{Z}^m,\mathbf{W}^m,\mathbf{S}^m} \sum_{m=1}^{v} \Bigg(\|(\mathbf{W}^m)^T\mathbf{X}^m - (\mathbf{W}^m)^T\mathbf{X}^m\mathbf{S}^m\|_F^2 + \gamma\|\mathbf{W}^m\|_{2,1} \\ &\quad + \rho \sum_{i,j=1}^{n} \|\phi^m(X_i^m) - \phi^m(X_j^m)\|_2^2 Z_{ij}^m \Bigg) \\ &\text{s.t. } (\mathbf{W}^m)^T\mathbf{X}^m(\mathbf{X}^m)^T\mathbf{W}^m = I_k, \\ &\quad \text{diag}(\mathbf{Z}^m) = 0, (\mathbf{z}_i^m)^T\mathbf{1} = 1, \mathbf{Z}^m \geq 0.\end{aligned} \tag{9}$$

To fully exploit the complementary strengths of global and local graph structures, we introduce a graph structure alignment mechanism. Specifically, we adopt a Frobenius norm-based graph discrepancy term that encourages the consistency between the global and local graphs during optimization. Even when these graphs are constructed from different perspectives or feature spaces, such as the projected space and a kernel-induced space, this mechanism ensures structural alignment, thereby enhancing the discriminative power and coherence of the learned graph.

Our design is inspired by the TAML framework proposed in [42], the term $\|\mathcal{X} - \mathcal{Y}\|_F^2$ is employed to enforce consistency between the coefficient tensor $\mathcal{X}$, which captures global self-representation structures, and the similarity tensor $\mathcal{Y}$, which models local geometric relationships via adaptive graphs. Although $\mathcal{X}$ and $\mathcal{Y}$ originate from distinct modeling approaches, they fundamentally encode pairwise sample affinities and share a common semantic structure based on sample indices, which makes direct alignment meaningful. Furthermore, the Frobenius norm is computationally efficient and differentiable, which makes it tractable during optimization.

The corresponding formulation is presented as follows.

$$\begin{aligned}&\min_{\mathbf{Z}^m,\mathbf{W}^m,\mathbf{S}^m} \sum_{m=1}^{v} \Bigg(\|(\mathbf{W}^m)^T\mathbf{X}^m - (\mathbf{W}^m)^T\mathbf{X}^m\mathbf{S}^m\|_F^2 + \gamma\|\mathbf{W}^m\|_{2,1} \\ &\quad + \rho \sum_{i,j=1}^{n} \|\phi^m(X_i^m) - \phi^m(X_j^m)\|_2^2 Z_{ij}^m \Bigg) + \frac{\beta}{2}\|\mathcal{S} - \mathcal{Z}\|_F^2 \\ &\text{s.t. } (\mathbf{W}^m)^T\mathbf{X}^m(\mathbf{X}^m)^T\mathbf{W}^m = I_k, \mathcal{Z} = \Phi(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(m)}), \\ &\quad \text{diag}(\mathbf{Z}^m) = 0, (\mathbf{z}_i^m)^T\mathbf{1} = 1, \mathbf{Z}^m \geq 0, \\ &\quad \mathcal{S} = \Phi(\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \ldots, \mathbf{S}^{(v)}),\end{aligned} \tag{10}$$

where the function $\Phi(\cdot)$ stacks multiple view-specific representations $\mathbf{Z}^{(m)}$ or $\mathbf{S}^{(m)}$ into a third-order tensor $\mathbf{Z}$ or $\mathbf{S}$.

3.2.2. Low-rank constraint regularization

Although conventional regularization strategies effectively constrain the representation within each individual view, they often fall short in capturing the intrinsic inter-view correlations and high-order dependencies that are essential to multi-view learning. To address this limitation, we leverage recent advances in low-rank tensor learning [43–46] and incorporate a tensor nuclear norm regularization based on tensor singular value decomposition (t-SVD). This constraint is applied to the subspace similarity representations, which facilitates the extraction of complementary information embedded in the high-order interactions across

multiple views. Specifically, the proposed model is formulated as

$$\begin{aligned} &\min \|\mathcal{S}\|_{\otimes} \\ &\text{s.t. } \mathcal{S} = \Phi(\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \ldots, \mathbf{S}^{(v)}), \end{aligned} \tag{11}$$

where the similarity matrices from each view, $\mathbf{S}^1, \mathbf{S}^2, \ldots, \mathbf{S}^v$ are aggregated into a third-order tensor $\mathbf{S} \in \mathbb{R}^{n\times n\times v}$. The t-SVD-based tensor nuclear norm $\|\mathcal{S}\|_{\otimes}$ imposes a low-rank structure on $\mathbf{S}$, which encourages compactness while preserving latent structural dependencies shared across different views.

Remark 1. Fig. 2 illustrates the computational pipeline of the t-SVD-based tensor low-rank approximation used in our model. Given multiple view-specific self-representation matrices $S^{(1)}, \ldots, S^{(V)}$, they are first stacked into a third-order tensor $\mathcal{S} \in \mathbb{R}^{n\times n\times V}$. This tensor is then rotated (or transformed) into $\tilde{S}$ for better alignment along the third (view) mode.

Next, the t-SVD (tensor singular value decomposition) is applied:

$$\tilde{\mathcal{S}} = \mathcal{U} * \mathcal{C} * \mathcal{V}^{\top},$$

where $\mathcal{U}$ and $\mathcal{V}$ are orthogonal tensors, and $\mathcal{C}$ is an f-diagonal core tensor containing the tubal singular values.

To promote low-rank structure and suppress noise, a *tubal-shrinkage* operator is applied to $\mathcal{C}$ in the Fourier domain, which yields a shrunk core tensor $\tilde{\mathcal{C}}$. The updated tensor is reconstructed via the t-product:

$$\tilde{\mathcal{S}} \leftarrow \mathcal{U} * \tilde{\mathcal{C}} * \mathcal{V}^{\top}.$$

Finally, the inverse rotation transforms $\tilde{\mathcal{S}}$ back to the original tensor space for downstream clustering or selection. This process enables effective multi-view structural modeling with shared low-rank priors.

3.2.3. Consensus clustering

In multi-view learning, each view offers a distinct yet complementary representation of the underlying data. To effectively integrate these heterogeneous sources, consensus clustering aims to identify a unified clustering assignment that reflects the shared structural patterns across all views. This is accomplished by learning a consensus indicator matrix that approximates the aggregated similarity matrices derived from all views. Formally, the consensus clustering objective is defined as

$$\begin{aligned} &\min_{\mathbf{S}^m} \left\| \sum_{m=1}^{v} \mathbf{S}^m - \mathbf{V}\mathbf{V}^{\top} \right\|_F^2 \\ &\text{s.t. } \mathbf{V}^{\top}\mathbf{V} = \mathbf{I}_k, \end{aligned} \tag{12}$$

where $\mathbf{V} \in \mathbb{R}^{n\times k}$ denotes the consensus cluster assignment matrix, and k is the predefined number of clusters.

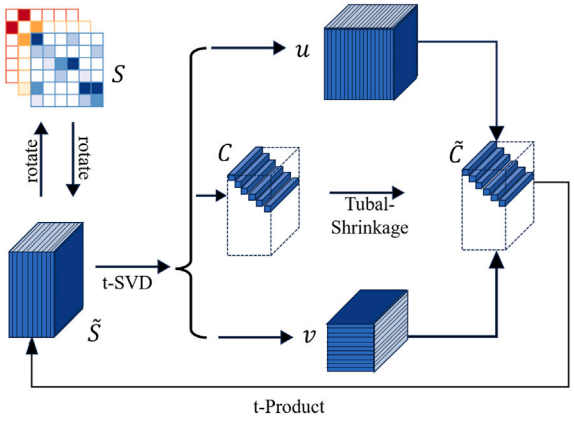


Fig. 2. The Flowchart of t-SVD-MSC. $\mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(V)}$, into a tensor $\mathcal{S}$, and then rotates to $\tilde{\mathcal{S}}$; the $\tilde{\mathcal{S}}$ will be updated by using t-SVD based tensor multi-rank minimization.

3.2.4. Overall objective function

By integrating the above three components, we obtain the final objective function.

$$\begin{aligned} &\min_{\mathbf{Z}^m, \mathbf{W}^m, \mathbf{S}^m, \mathbf{V}} \alpha\|\mathcal{S}\|_{\otimes} + \frac{\beta}{2}\|\mathcal{S} - \mathcal{Z}\|_F^2 \\ &\quad + \sum_{m=1}^{v} \left(\|(\mathbf{W}^m)^T\mathbf{X}^m - (\mathbf{W}^m)^T\mathbf{X}^m\mathbf{S}^m\|_F^2 + \gamma\|\mathbf{W}^m\|_{2,1} \right. \\ &\quad \left. + \rho \sum_{i,j=1}^{n} \|\phi^m(X_i^m) - \phi^m(X_j^m)\|_2^2 Z_{ij}^m \right) \\ &\quad + \left\| \sum_{m=1}^{v} \mathbf{S}^m - \mathbf{V}\mathbf{V}^T \right\|_F^2 \\ &\text{s.t. } (\mathbf{W}^m)^T\mathbf{X}^m(\mathbf{X}^m)^T\mathbf{W}^m = I_k, \mathbf{V}^T\mathbf{V} = I_k, \\ &\quad \mathcal{Z} = \Phi(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(m)}), \mathrm{diag}(\mathbf{Z}^m) = 0, \\ &\quad (\mathbf{z}_i^m)^T\mathbf{1} = 1, \mathbf{Z}^m \geq 0, \\ &\quad \mathcal{S} = \Phi(\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \ldots, \mathbf{S}^{(v)}). \end{aligned} \tag{13}$$

3.3. Optimization

In the presence of multiple interdependent variables, obtaining a closed-form solution to the constrained optimization problem in Eq. (13) becomes analytically intractable. To address this challenge, we reformulate the original objective into a sequence of more manageable sub-problems and solve them iteratively using the Alternating Direction Method of Multipliers (ALM-ADM) within the Augmented Lagrangian framework.

To facilitate the decoupling of the objective function and enable efficient optimization, we introduce an auxiliary tensor variable $\mathcal{J} \in \mathbb{R}^{n\times n\times v}$. The reformulated objective is expressed as follows.

$$\begin{aligned} &\min_{\mathbf{Z}^m, \mathbf{W}^m, \mathbf{S}^m, \mathbf{V}} \alpha\|\mathcal{J}\|_{\otimes} + \frac{\beta}{2}\|\mathcal{S} - \mathcal{Z}\|_F^2 \\ &\quad + \sum_{m=1}^{v} \left(\|(\mathbf{W}^m)^T\mathbf{X}^m - (\mathbf{W}^m)^T\mathbf{X}^m\mathbf{S}^m\|_F^2 + \gamma\|\mathbf{W}^m\|_{2,1} \right. \\ &\quad \left. + \rho \sum_{i,j=1}^{n} \|\phi^m(X_i^m) - \phi^m(X_j^m)\|_2^2 Z_{ij}^m \right) \\ &\quad + \left\| \sum_{m=1}^{v} \mathbf{S}^m - \mathbf{V}\mathbf{V}^T \right\|_F^2 \\ &\text{s.t. } (\mathbf{W}^m)^T\mathbf{X}^m(\mathbf{X}^m)^T\mathbf{W}^m = I_k, \mathbf{V}^T\mathbf{V} = I_k, \\ &\quad \mathcal{Z} = \Phi(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(m)}), \mathrm{diag}(\mathbf{Z}^m) = 0, \\ &\quad (\mathbf{z}_i^m)^T\mathbf{1} = 1, \mathbf{Z}^m \geq 0, \\ &\quad \mathcal{S} = \Phi(\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \ldots, \mathbf{S}^{(v)}), \mathcal{J} = \mathcal{S}. \end{aligned} \tag{14}$$

The corresponding augmented Lagrangian function is then formulated as

$$\begin{aligned} &\min_{\mathbf{Z}^m, \mathbf{W}^m, \mathbf{S}^m, \mathbf{V}} \alpha\|\mathcal{J}\|_{\otimes} + \frac{u}{2}\left\|\mathcal{J} - \left(\mathcal{S} + \frac{\mathcal{M}}{u}\right)\right\|_F^2 \\ &\quad + \sum_{m=1}^{v} \left(\|(\mathbf{W}^m)^T\mathbf{X}^m - (\mathbf{W}^m)^T\mathbf{X}^m\mathbf{S}^m\|_F^2 + \gamma\|\mathbf{W}^m\|_{2,1} \right. \\ &\quad \left. + \rho \sum_{i,j=1}^{n} \|\phi^m(X_i^m) - \phi^m(X_j^m)\|_2^2 Z_{ij}^m \right) \\ &\quad + \frac{\beta}{2}\|\mathcal{S} - \mathcal{Z}\|_F^2 + \left\| \sum_{m=1}^{v} \mathbf{S}^m - \mathbf{V}\mathbf{V}^T \right\|_F^2 \\ &\text{s.t. } (\mathbf{W}^m)^T\mathbf{X}^m(\mathbf{X}^m)^T\mathbf{W}^m = I_k, \mathbf{V}^T\mathbf{V} = I_k, \\ &\quad \mathcal{Z} = \Phi(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(m)}), \mathrm{diag}(\mathbf{Z}^m) = 0, \\ &\quad (\mathbf{z}_i^m)^T\mathbf{1} = 1, \mathbf{Z}^m \geq 0, \\ &\quad \mathcal{S} = \Phi(\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \ldots, \mathbf{S}^{(v)}) \end{aligned} \tag{15}$$

where $\mathcal{M} \in \mathbb{R}^{n\times n\times v}$ denotes the Lagrange multiplier tensor and u is a positive penalty parameter. Following the the alternative minimization strategy, we decompose the problem in Eq. (15) into a set of subproblems, each of which optimizes single variable while keeping the others fixed. The complete optimization procedure is detailed in the subsequent sections.

1) $\mathbf{Z}^m$-Subproblem: When all other variables are held constant, the subproblem with respect to $\mathbf{Z}^m$ is expressed as

$$\begin{aligned}&\min_{\mathbf{Z}^m} \frac{\beta}{2}\|S-\mathcal{Z}\|_F^2+\rho\sum_{i,j=1}^{n}\|\phi^m(X_i^m)-\phi^m(X_j^m)\|_2^2 Z_{ij}^m\\&\text{s.t. } \mathcal{Z}=\Phi(\mathbf{Z}^{(1)},\mathbf{Z}^{(2)},\ldots,\mathbf{Z}^{(m)}),\operatorname{diag}(\mathbf{Z}^m)=0,\\&\quad(\mathbf{z}_i^m)^T\mathbf{1}=1,\mathbf{Z}^m\geq 0.\end{aligned}\tag{16}$$

Let $d_{ij}^x = \|\mathbf{x}_i-\mathbf{x}_j\|^2 = x_i^Tx_i + x_j^Tx_j - 2x_i^Tx_j$ denote the squared Euclidean distance between samples x_i and x_j in the original feature space. This pairwise distance can be compactly represented in matrix form as $\mathbf{D}^x = \operatorname{Diag}(\mathbf{X}\mathbf{X}^T)\mathbf{1}\mathbf{1}^T + \mathbf{1}\mathbf{1}^T\operatorname{Diag}(\mathbf{X}\mathbf{X}^T) - 2\mathbf{X}\mathbf{X}^T$, where $\operatorname{Diag}(\mathbf{X}\mathbf{X}^T)$ denotes a diagonal matrix containing the diagonal elements of the Gram matrix $\mathbf{X}\mathbf{X}^T$ [47]. Extending this formulation to the reproducing kernel Hilbert space (RKHS), the pairwise distance between samples in the m-th view is defined as $(d_{ij}^m)^x = \|\phi^m(X_i^m)-\phi^m(X_j^m)\|^2 = (\phi^m(x_i^m))^T\phi^m(x_i^m) + (\phi^m(x_j^m))^T\phi^m(x_j^m) - 2(\phi^m(x_i^m))^T\phi^m(x_j^m)$, where $\phi^m(.)$ denotes the implicit nonlinear mapping associated with the kernel function of the view. Let $K^{(m)} = \phi(X^{(m)})^T\phi(X^{(m)})$ denote the corresponding kernel matrix. The resulting distance matrix in RKHS is then given by

$$(\mathbf{D}^m)^x = \operatorname{Diag}(\mathbf{K}^m)\mathbf{1}\mathbf{1}^T + \mathbf{1}\mathbf{1}^T\operatorname{Diag}(\mathbf{K}^m) - 2\mathbf{K}^m.\tag{17}$$

Given that $\Phi_{(m)}^{-1}(\mathcal{Z}) = \mathbf{Z}^{(m)}$ and $\Phi_{(m)}^{-1}(S) = \mathbf{S}^{(m)}$, Eq. (16) can be equivalently reformulated in a trace-based representation as follows:

$$\begin{aligned}&\min_{\mathbf{Z}^m}\frac{\beta}{2}\|\mathbf{S}^m-\mathbf{Z}^m\|_F^2+\rho\operatorname{Tr}((\mathbf{Z}^m)^T(\mathbf{D}^m)^x)\\&\text{s.t. } \operatorname{diag}(\mathbf{Z}^m)=0,(\mathbf{z}_i^m)^T\mathbf{1}=1,\mathbf{Z}^m\geq 0.\end{aligned}\tag{18}$$

To derive the analytical solution to the optimization problem defined in Eq. (18), we set the derivative with respect to $\mathbf{Z}^m$ to zero, which yields

$$\mathbf{Z}^m = \mathbf{S}^m - \frac{\rho\mathbf{D}^m}{\beta}.\tag{19}$$

2) $\mathbf{W}^m$-Subproblem: By omitting irrelevant components, the subproblem corresponding to the projection matrix $\mathbf{W}^m$ for the view can be reformulated as

$$\begin{aligned}&\min_{\mathbf{W}^m}\|(\mathbf{W}^m)^T\mathbf{X}^m-(\mathbf{W}^m)^T\mathbf{X}^m\mathbf{S}^m\|_F^2+\gamma\|\mathbf{W}^m\|_{2,1}\\&\text{s.t. } (\mathbf{W}^m)^T\mathbf{X}^m(\mathbf{X}^m)^T\mathbf{W}^m=I_k.\end{aligned}\tag{20}$$

To address the non-smooth nature of the $l_{2,1}$ regularization, we incorporate a diagonal weighting matrix $\mathbf{D}_m$, whose diagonal elements are defined as

$$\mathbf{D}_{m[i,i]} = \frac{1}{\max\left(2\left\|\mathbf{W}_i^m\right\|_2,\epsilon\right)},\tag{21}$$

to prevent numerical instability caused by the singularity condition $\left\|\mathbf{W}_i^m\right\|_2 = 0$, a small positive constant ϵ is introduced as a regularization term. By defining $\mathbf{U}^m = (\mathbf{I}^m-\mathbf{S}^m)(\mathbf{I}^m-\mathbf{S}^m)^\top$ and incorporating Eq. (21), the subproblem for optimizing $\mathbf{W}^m$ can be reformulated as follows

$$\begin{aligned}&\min_{\mathbf{W}^m}\operatorname{Tr}\left[(\mathbf{W}^m)^T\left(\mathbf{X}^m\mathbf{U}^m(\mathbf{X}^m)^T+\gamma\mathbf{D}_m\right)\mathbf{W}^m\right]\\&\text{s.t. } (\mathbf{W}^m)^T\mathbf{X}^m(\mathbf{X}^m)^T\mathbf{W}^m=I_k.\end{aligned}\tag{22}$$

The optimal $\mathbf{W}^m$ can be obtained by solving the following generalized eigen-problem

$$\left(\mathbf{X}^m\mathbf{U}^m(\mathbf{X}^m)^T+\gamma\mathbf{D}^m\right)\mathbf{W}^m = \Lambda^m\mathbf{X}^m(\mathbf{X}^m)^T\mathbf{W}^m,\tag{23}$$

where Λ^m is a diagonal matrix whose entries correspond to the associated eigenvalues. It should be noted, Eq. (23) requires $\mathbf{X}^m(\mathbf{X}^m)^T$ is non-singular. Furthermore, the computational complexity of this solution is $O\left(d^3+nd^3\right)$, which renders it inefficient for high-dimensional data. To circumvent these challenges, we adopt the method proposed in [48], which approximates the optimal $\mathbf{W}^m$ by solving the following problem

$$\min_{\mathbf{W}^m}\|\mathbf{Y}^m-(\mathbf{X}^m)^T\mathbf{H}^m\|_F^2+\gamma\|\mathbf{W}^m\|_{2,1},\tag{24}$$

where $\mathbf{Y}^m$ consists of the eigenvectors associated with the k smallest eigenvalues derived from the spectral decomposition $\mathbf{U}^m\mathbf{Y}^m = \Lambda^m\mathbf{Y}^m$. Using the diagonal matrix defined in Eq. (21) and the Iterative Reweighted Least-Squares (IRLS) [49] algorithm, the optimal $\mathbf{W}^m$ can be obtained by

$$\mathbf{W}^m = \left(\mathbf{X}^m(\mathbf{X}^m)^T+\gamma\mathbf{D}^m\right)^{-1}\mathbf{X}^m\mathbf{Y}^m.\tag{25}$$

3) $\mathbf{S}^m$-Subproblem: We define $\Phi_{(m)}^{-1}(\mathcal{J}) = \mathbf{J}^{(m)}$, $\Phi_{(m)}^{-1}(S) = \mathbf{S}^{(m)}$ and $\Phi_{(m)}^{-1}(\mathcal{M}) = \mathbf{M}^{(m)}$. With all other variables fixed, the optimization subproblem concerning the affinity matrix $\mathbf{S}^m$ for the m-th view is formulated as follows:

$$\begin{aligned}&\min_{\mathbf{S}^m}\frac{u}{2}\left\|\mathbf{J}^m-\left(\mathbf{S}^m+\frac{\mathbf{M}^m}{u}\right)\right\|_F^2+\frac{\beta}{2}\|\mathbf{S}^m-\mathbf{Z}^m\|_F^2\\&\quad+\|(\mathbf{W}^m)^T\mathbf{X}^m-(\mathbf{W}^m)^T\mathbf{X}^m\mathbf{S}^m\|_F^2\\&\quad+\left\|\sum_{m=1}^{v}\mathbf{S}^m-\mathbf{V}\mathbf{V}^T\right\|_F^2.\end{aligned}\tag{26}$$

To derive a closed-form solution, we introduce the notation $\mathbf{Q}^m = (\mathbf{X}^m)^T\mathbf{W}^m(\mathbf{W}^m)^T\mathbf{X}^m$. By setting the gradient of the objective in Eq. (26) with respect to $\mathbf{S}^m$ to zero, we obtain the following analytical expression:

$$\begin{aligned}\mathbf{S}^m = (2\mathbf{Q}^m+(2+u+\beta)I_n)^{-1}\Bigg(&2\left(\mathbf{Q}^m+\mathbf{V}\mathbf{V}^T-\sum_{v\neq m}^{V}\mathbf{S}^v\right)\\&+\beta\mathbf{Z}^m-u\mathbf{J}^m-\mathbf{M}^m\Bigg).\end{aligned}\tag{27}$$

4) $\mathbf{V}$-Subproblem: By means of elementary algebraic manipulations, the optimization subproblem with respect to V can be reformulated as

$$\begin{aligned}&\min_{\mathbf{V}}\operatorname{Tr}\left[\mathbf{V}^\top\left(\mathbf{I}_n-2\sum_{m=1}^{v}(\mathbf{S}^m)^T\right)\mathbf{V}\right]\\&\text{s.t. } \mathbf{V}^T\mathbf{V}=\mathbf{I}_k.\end{aligned}\tag{28}$$

The optimal solution for V is obtained efficiently through SVD, which guarantees an orthonormal basis satisfying the orthogonality constraint.

5) $\mathcal{J}$-Subproblem: By excluding terms that do not affect the current optimization, the subproblem for the auxiliary tensor variable $\mathcal{J}$ simplifies to

$$\min_{\mathcal{J}}\alpha\|\mathcal{J}\|_\otimes+\frac{u}{2}\left\|\mathcal{J}-\left(\mathcal{S}+\frac{\mathcal{M}}{u}\right)\right\|_F^2\tag{29}$$

Letting $\mathcal{X} = \mathcal{S}+\frac{\mathcal{M}}{u}$, Eq. (29) can be efficiently addressed by applying Theorem 1 [50], which states:

Theorem 1. *Consider third-order tensors $\mathcal{J}, \mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and a positive scalar $\alpha > 0$. The problem*

$$\min_{\mathcal{J}} \alpha \|\mathcal{J}\|_{\otimes} + \frac{1}{2}\|\mathcal{J} - \mathcal{X}\|_F^2 \tag{30}$$

admits a closed-form solution via the tensor tubal-shrinkage operator given by

$$\mathcal{J} = C_{N_3\tau}(\mathcal{X}) = \mathcal{U} * C_{N_3\tau}(\mathcal{O}) * \mathcal{V}^T \tag{31}$$

where $\mathcal{X} = \mathcal{U} * \mathcal{O} * \mathcal{V}^T$ and $C_{N_3\tau}(\mathcal{O}) = \mathcal{O} * \mathcal{Y}$. Here, $\mathcal{Y} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is an f-diagonal tensor, with its diagonal elements of $\mathcal{Y}$ is defined as

$$\mathcal{Y}_f(i,i,j) = (1 - (N_3\tau / \mathcal{O}(i,i,j)))_+. \tag{32}$$

The principal stages of the proposed method's optimization process are delineated in Algorithm 1.

Algorithm 1 GDFS.[51].

Input: Data matrix $\mathbf{X}_m$ and parameters $\alpha, \beta, \gamma, \rho$.
Initialize: $\mathbf{S}_m$ by WPKN algorithm [51], $\mathbf{W}_m = \text{rand}(d_m, k)$, $\mathbf{Z} = \text{rand}(n, n)$, $\mathbf{V} = \text{rand}(n, k)$.
repeat
  Update $\mathbf{Z}_m$ by Eq. (19);
  Update $\mathbf{W}_m$ by Eq. (25);
  Update $\mathbf{S}_m$ by Eq. (27);
  Update $\mathbf{V}$ by Eq. (28);
  Update $\mathcal{J}$ by Eq. (29);
until $\frac{\text{obj}^{(t-1)} - \text{obj}^{(t)}}{\text{obj}^{(t)}} < eps$;
Output: Feature weight matrix $\mathbf{W}_m$.
Feature selection: Arrange all features of the multi-view data in descending order according to $\left\|\mathbf{W}_{m[:,i]}\right\|_2$ and select the k top-ranked ones.

3.4. Analysis

Convergence: Given that Algorithm 1 contains multiple iterative updates, it is important to discuss its convergence behavior. Theoretical results have shown that the Alternating Direction Method (ADM) converges under certain conditions when only two variables are updated alternately [52]. However, when the number of variables increases to N ($N \geq 3$), proving convergence becomes significantly more challenging [53]. In our case, the algorithm simultaneously updates several variables, which include $\mathbf{Z}^m$, $\mathbf{W}^m$, $\mathbf{S}^m$, $\mathbf{V}$, and $\mathcal{J}$, which complicates the derivation of strict theoretical guarantees.

Nevertheless, GDFS mitigates the risk of reinforcing noisy or suboptimal clustering partitions through a joint optimization framework that ensures mutual enhancement between the consensus clustering matrix $\mathbf{V}$ and feature selection. Specifically, the alternating optimization strategy guarantees that each update of $\mathbf{V}$ is conditioned on the most recent feature selection result, and vice versa, which ensures progressive refinement toward a more robust consensus. Additionally, the imposed low-rank constraint on the global graph tensor $\mathcal{J}$ helps filter out cross-view inconsistencies, which in turn stabilizes the pseudo-labels and improves convergence behavior.

Moreover, each subproblem in our framework has an optimal solution, which ensures the reduction of the objective function and further improves the overall performance. Empirically, as shown in Section 4.6, our algorithm consistently converges within a few iterations on multiple real-world datasets.

Complexity: The computational burden of the method is primarily dictated by the iterative updates of these five variables. The update of $\mathbf{Z}^m$ requires matrix inversion, leading to a complexity of $o(n^3)$. Updating $\mathbf{W}^m$ involves solving an eigenvalue decomposition combined with sparse feature selection, incurring a complexity of $O(kn^2 + d_m^3)$. Similarly, $\mathbf{S}^m$ demands matrix inversion, also resulting in $O(n^3)$ complexity. SVD for updating $\mathbf{V}$ contributes $O(kn^2)$ complexity. The update of $\mathcal{J}$ requires 3D Fast Fourier Transforms (FFT) and inverse FFTs on an $n \times v \times n$ tensor, coupled with n SVD computations on $n \times v$ matrices within the Fourier domain, which culminates in a complexity of $O(2n^2 v \log(n))$. By aggregating across all views, the total per-iteration complexity is $O(vn^3 + \sum_{m=1}^{v} d_m^3 + 2n^2 v \log(n))$.

Discussion: GDFS embodies several notable advantages. By integrating both local and global graph structures, it captures overall distribution and the neighborhood information of data, thereby facilitating richer and more nuanced representations of sample interrelationships. The adoption of kernel mappings enables the exploration of nonlinear local structures, while global graph constraints preserve structural consistency across different views. The introduction of a graph discrepancy term effectively aligns heterogeneous graphs, which fosters a robust and unified feature selection mechanism. Additionally, the imposition of a low-rank tensor constraint on the stacked global graphs enhances cross-view correlation modeling, which is particularly advantageous when handling heterogeneous data sources. To the best of our knowledge, this method represents one of the first attempts to synergize graph discrepancy learning with consensus pseudo-label guidance in multi-view unsupervised feature selection.

4. Experiments

4.1. Datasets and compared methods

In this section, we assess the efficacy of the proposed GDFS method across a suite of real-world multi-view datasets. Detailed characteristics of these datasets are summarized in Table 3.

To comprehensively evaluate GDFS, we carried out extensive comparative experiments with both classical and state-of-the-art multi-view feature selection algorithms. A brief overview of these competing approaches is provided below. In addition, to further verify the effectiveness of our method in capturing cross-view correlations, we also include SLNMF (Soft-label guided Non-negative Matrix Factorization for Unsupervised Feature Selection) [54], a recent single-view feature selection method. Since SLNMF is not designed for multi-view data, we follow a common practice of concatenating all views before applying feature selection. The performance is evaluated on the same datasets as our method to ensure fairness.

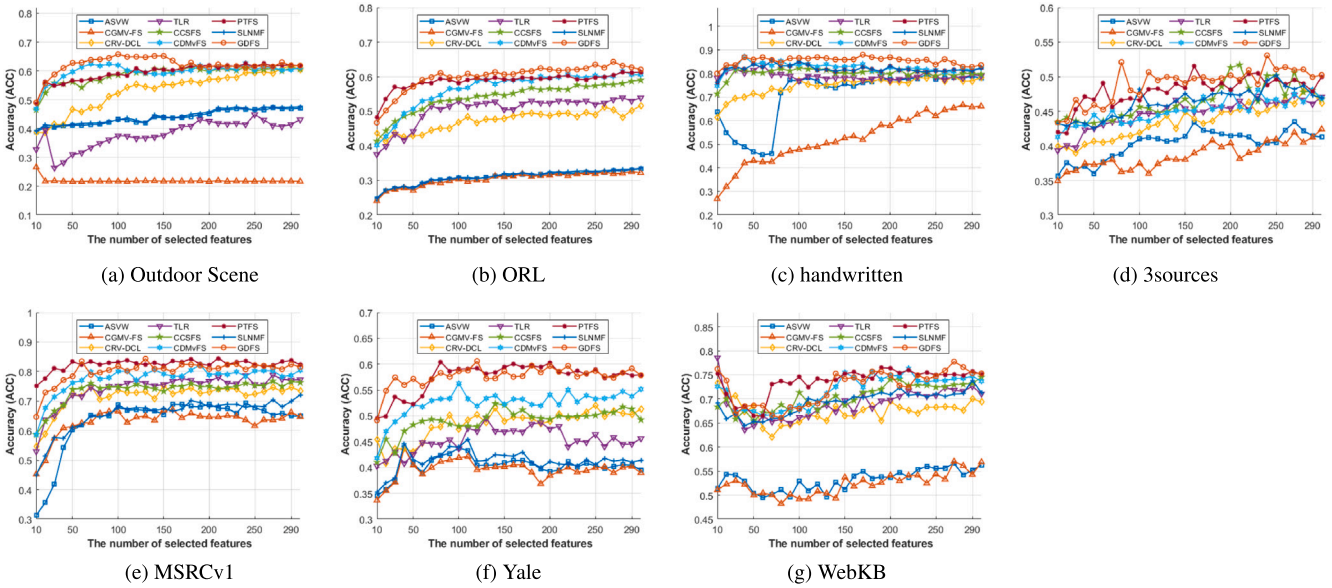
(1) ASVW [55] first learns an underlying consensus graph and then utilizes this consensus graph to ensure that the transformed data preserves local structures.
(2) CGMV-FS [15] employs non-negative matrix factorization alongside consensus learning to extract informative features spanning multiple views.
(3) CRV-DCL [56] maps the original data into a shared label space, decomposed into consensus and diversity components, to effectively identify discriminative features.
(4) TLR [40] integrates multiple graphs into a tensor-based framework regulated by low-rank constraints, which capture high-order inter-view dependencies.

Table 3
Details of datasets.

Datasets	Class	View	Samples	Features
Outdoor Scene	8	4	2688	512,432,256,48
ORL	40	3	400	4096,3304,6750
handwritten	10	6	2000	76,216,64,240,47,6
3Sources	6	3	169	3560,3631,3068
MSRCV1	7	5	210	24,576,512,254,256
Yale	15	3	165	4096,3304,6750
WebKB	4	3	203	1703,230,230

Table 4
Clustering performance on different datasets.

Dataset	Metric	ASVW	CGMV-FS	CRV-DCL	TLR	CCSFS	CDMvFS	PTFS	SLNMF	GDFS
Outdoor Scene	ACC	47.22 ± 1.47	26.61 ± 0.83	61.71 ± 3.64	44.83 ± 2.94	61.45 ± 3.56	62.60 ± 4.38	62.53 ± 4.26	47.52 ± 2.32	**65.79 ± 3.20**
	NMI	39.91 ± 1.17	11.74 ± 0.46	49.14 ± 0.43	38.11 ± 1.12	51.88 ± 1.91	52.19 ± 0.63	53.70 ± 1.23	40.24 ± 0.90	**54.23 ± 1.45**
	ARI	28.53 ± 0.52	6.58 ± 0.30	40.57 ± 0.71	25.79 ± 1.20	41.05 ± 1.73	41.74 ± 1.68	43.11 ± 2.49	28.68 ± 0.81	**45.21 ± 2.05**
	F-score	38.10 ± 0.47	19.37 ± 0.70	48.39 ± 0.58	35.75 ± 1.00	48.70 ± 1.47	49.38 ± 0.89	50.60 ± 1.93	38.18 ± 0.70	**52.30 ± 1.75**
	Precision	36.25 ± 0.77	18.00 ± 0.30	46.79 ± 2.51	33.85 ± 1.34	47.65 ± 1.75	48.57 ± 3.02	48.96 ± 3.44	36.69 ± 0.91	**51.37 ± 2.25**
ORL	ACC	33.33 ± 1.53	32.56 ± 1.36	55.08 ± 2.72	53.94 ± 3.26	58.98 ± 2.85	61.59 ± 3.01	61.31 ± 3.46	33.55 ± 1.53	**64.30 ± 2.50**
	NMI	55.40 ± 1.23	54.88 ± 1.49	74.02 ± 1.79	73.66 ± 1.38	76.88 ± 1.82	78.50 ± 1.87	78.29 ± 2.21	56.04 ± 1.50	**80.33 ± 1.71**
	ARI	14.74 ± 1.00	14.43 ± 1.11	40.68 ± 2.98	39.32 ± 2.72	45.91 ± 3.13	49.09 ± 3.42	48.57 ± 4.33	15.23 ± 1.06	**52.16 ± 3.29**
	F-score	17.10 ± 0.90	16.82 ± 1.01	42.22 ± 2.90	40.94 ± 2.63	47.29 ± 3.03	50.37 ± 3.32	49.89 ± 4.20	17.57 ± 0.98	**53.36 ± 3.21**
	Precision	14.04 ± 1.01	13.64 ± 1.22	36.70 ± 2.92	35.16 ± 4.10	41.90 ± 3.38	45.01 ± 3.49	43.99 ± 4.49	14.48 ± 1.30	**48.00 ± 3.29**
handwritten	ACC	79.60 ± 7.86	66.69 ± 4.77	78.78 ± 6.14	83.77 ± 7.28	82.31 ± 5.56	87.10 ± 5.79	85.44 ± 7.38	80.45 ± 6.60	**88.01 ± 6.14**
	NMI	77.76 ± 3.80	66.57 ± 3.09	78.69 ± 2.92	82.89 ± 4.16	81.91 ± 4.19	82.88 ± 2.63	84.24 ± 2.43	78.22 ± 3.87	**86.17 ± 3.43**
	ARI	71.26 ± 6.36	55.02 ± 4.71	71.90 ± 4.63	77.52 ± 7.22	76.06 ± 7.07	78.62 ± 5.52	78.77 ± 5.88	72.31 ± 7.10	**82.46 ± 6.60**
	F-score	74.27 ± 5.62	59.73 ± 4.15	74.86 ± 4.10	79.85 ± 6.42	78.58 ± 6.27	80.81 ± 4.91	80.97 ± 5.21	75.20 ± 6.28	**84.29 ± 5.85**
	Precision	71.30 ± 7.02	56.94 ± 4.83	71.31 ± 5.87	77.29 ± 8.06	75.17 ± 8.12	79.30 ± 6.40	78.51 ± 7.17	72.38 ± 8.14	**81.65 ± 8.25**
3sources	ACC	43.46 ± 5.79	42.41 ± 5.79	47.96 ± 8.62	47.34 ± 6.70	51.70 ± 8.04	48.14 ± 4.81	51.54 ± 8.86	50.21 ± 8.19	**53.08 ± 7.71**
	NMI	19.85 ± 6.39	17.60 ± 4.83	28.95 ± 11.08	29.89 ± 8.43	28.97 ± 8.90	25.85 ± 9.51	**33.72 ± 7.41**	25.64 ± 7.22	31.33 ± 6.51
	ARI	7.39 ± 8.59	5.94 ± 6.52	16.94 ± 11.89	16.00 ± 9.94	21.01 ± 11.61	15.63 ± 17.17	21.20 ± 11.81	15.93 ± 13.91	**22.50 ± 13.89**
	F-score	37.25 ± 4.75	36.99 ± 3.09	42.50 ± 6.72	41.27 ± 6.42	45.46 ± 6.53	43.96 ± 9.56	45.91 ± 6.43	43.04 ± 8.26	**46.01 ± 8.49**
	Precision	27.22 ± 5.05	26.36 ± 3.59	32.85 ± 6.91	32.17 ± 5.66	34.96 ± 7.13	31.58 ± 9.40	35.72 ± 7.74	31.89 ± 7.91	**36.74 ± 9.75**
MSRCV1	ACC	69.38 ± 6.27	67.02 ± 7.32	75.19 ± 5.28	78.81 ± 8.33	76.78 ± 6.79	82.26 ± 5.44	**84.51 ± 6.39**	77.81 ± 6.19	84.36 ± 7.93
	NMI	61.18 ± 3.66	58.37 ± 4.02	68.99 ± 6.68	73.17 ± 7.64	71.18 ± 5.60	75.39 ± 6.57	78.29 ± 5.58	71.82 ± 5.96	**78.72 ± 6.55**
	ARI	52.32 ± 4.88	48.86 ± 7.12	60.23 ± 7.03	66.87 ± 11.36	63.45 ± 8.23	67.94 ± 9.73	**73.08 ± 10.33**	65.07 ± 8.28	73.03 ± 9.09
	F-score	59.31 ± 4.10	56.37 ± 5.95	66.14 ± 8.09	71.79 ± 9.49	68.91 ± 6.81	72.70 ± 8.06	**76.98 ± 8.70**	70.29 ± 6.82	76.91 ± 7.71
	Precision	56.22 ± 4.53	53.66 ± 6.27	62.95 ± 6.95	68.43 ± 11.86	64.98 ± 6.47	69.45 ± 5.76	74.83 ± 10.49	66.20 ± 6.78	**74.97 ± 8.97**
Yale	ACC	43.82 ± 2.61	43.76 ± 2.44	52.00 ± 5.02	49.58 ± 4.35	52.38 ± 5.17	56.27 ± 5.64	60.40 ± 3.90	45.45 ± 3.28	**60.58 ± 6.77**
	NMI	49.65 ± 2.55	49.29 ± 1.70	60.47 ± 4.86	54.62 ± 3.18	56.75 ± 4.38	60.49 ± 5.21	**67.28 ± 5.10**	50.80 ± 2.42	66.04 ± 4.73
	ARI	23.47 ± 1.69	23.35 ± 1.71	36.12 ± 7.44	29.84 ± 3.66	32.70 ± 5.79	37.22 ± 5.48	**46.00 ± 7.47**	24.25 ± 1.54	44.89 ± 6.90
	F-score	28.78 ± 1.50	28.69 ± 1.58	41.06 ± 6.59	34.65 ± 3.18	37.22 ± 5.35	41.38 ± 6.39	**49.69 ± 6.83**	29.51 ± 1.31	48.53 ± 6.39
	Precision	25.33 ± 1.86	25.10 ± 1.60	33.02 ± 7.44	31.12 ± 4.11	33.88 ± 5.29	38.38 ± 5.44	44.95 ± 7.51	26.01 ± 1.86	**45.07 ± 6.78**
WebKB	ACC	56.60 ± 6.06	57.01 ± 6.51	73.30 ± 0.79	**78.60 ± 1.28**	74.52 ± 7.16	76.35 ± 8.30	76.49 ± 6.38	73.92 ± 5.32	77.78 ± 7.62
	NMI	15.83 ± 7.77	15.19 ± 6.42	41.37 ± 3.72	44.15 ± 4.56	48.21 ± 5.84	47.97 ± 5.87	**48.81 ± 3.67**	43.10 ± 4.83	48.34 ± 3.55
	ARI	13.20 ± 8.96	13.29 ± 7.07	44.66 ± 5.85	56.18 ± 2.15	53.66 ± 8.87	54.55 ± 7.07	**57.75 ± 7.88**	48.31 ± 8.29	56.60 ± 5.28
	F-score	54.97 ± 0.57	54.93 ± 0.58	65.30 ± 5.78	74.73 ± 1.46	71.67 ± 6.60	72.76 ± 5.28	75.12 ± 5.75	68.81 ± 4.53	**75.20 ± 6.30**
	Precision	46.33 ± 5.22	46.57 ± 4.28	69.75 ± 2.93	70.24 ± 3.49	74.13 ± 2.41	72.82 ± 2.47	73.73 ± 2.11	71.94 ± 2.09	**75.83 ± 3.42**



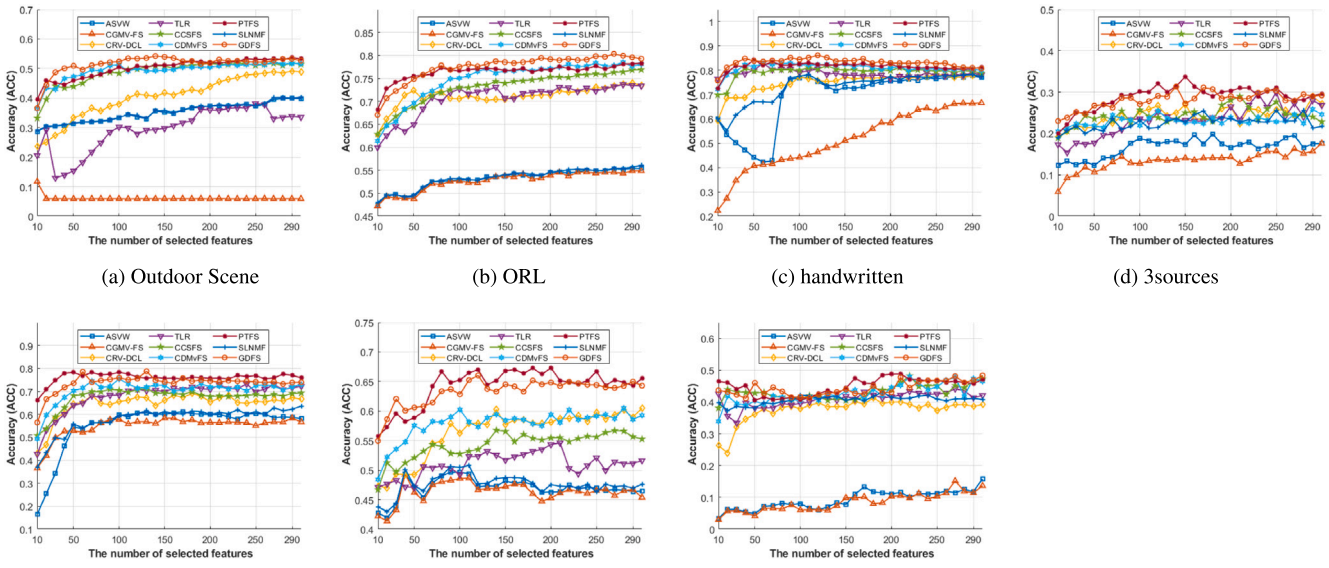
(a) Outdoor Scene (b) ORL (c) handwritten (d) 3sources

(e) MSRCv1 (f) Yale (g) WebKB

Fig. 3. The best ACC of different methods.

(5) CCSFS [57] leverages partition-level information to build a consensus label matrix, which enhances the discriminative capability of selected features.

(6) CDMvFS [28] produces multiple mutually exclusive graphs to strengthen inter-view complementarity, and couples graph learning with clustering through consistency measures.



(a) Outdoor Scene (b) ORL (c) handwritten (d) 3sources

(e) MSRCv1 (f) Yale (g) WebKB

Fig. 4. The best NMI of different methods.

Table 5
The paired t-test results of ACC of GDFS and comparison algorithms on all datasets.

Method	Outdoor scene		ORL		Handwritten		3sources		MSRCV1		Yale		WebKB	
	h	p	h	p	h	p	h	p	h	p	h	p	h	p
TLR	1	7.83 E-21	1	6.90 E-26	1	5.82 E-15	1	1.55 E-15	1	4.78 E-14	1	4.69 E-25	1	8.90 E-09
CCSFS	1	2.73 E-08	1	1.84 E-22	1	4.59 E-18	1	1.93 E-07	1	4.78 E-20	1	1.02 E-20	1	1.25 E-04
CDMvFS	1	4.59 E-07	1	1.71 E-10	1	8.12 E-11	1	1.01 E-11	1	1.26 E-08	1	6.85 E-17	1	3.89 E-02
PTFS	1	4.68 E-05	1	7.64 E-04	1	1.71 E-12	1	5.40 E-03	0	4.16 E-02	0	4.22 E-01	0	5.56 E-02

Table 6
The paired t-test results of NMI of GDFS and comparison algorithms on all datasets.

Method	Outdoor scene		ORL		Handwritten		3sources		MSRCV1		Yale		WebKB	
	h	p	h	p	h	p	h	p	h	p	h	p	h	p
TLR	1	1.21 E-17	1	5.80 E-24	1	2.36 E-14	1	1.94 E-11	1	1.31 E-08	1	5.00 E-27	1	7.72 E-09
CCSFS	1	4.49 E-09	1	2.05 E-21	1	6.74 E-13	1	9.55 E-11	1	1.09 E-16	1	2.30 E-24	1	4.20 E-03
CDMvFS	1	1.04 E-08	1	6.95 E-10	1	6.40 E-14	1	7.57 E-16	1	1.65 E-09	1	1.01 E-18	1	1.30 E-03
PTFS	1	1.94 E-03	1	2.18 E-04	1	1.21 E-09	0	5.65 E-02	0	5.70 E-02	0	5.11 E-02	0	1.81 E-01

Table 7
Comparison of metrics across noise levels in WebKB.

Noise α	ACC	NMI	ARI	F-score	Precision
0	77.78 ± 7.62	48.34 ± 3.55	56.60 ± 5.28	75.20 ± 6.30	75.83 ± 3.42
0.1	77.76 ± 5.47	48.41 ± 3.18	56.59 ± 5.11	75.09 ± 8.83	74.04 ± 2.23
0.3	76.28 ± 3.26	44.62 ± 3.72	53.33 ± 4.03	71.79 ± 2.61	71.36 ± 2.65
0.5	72.98 ± 2.77	42.29 ± 3.49	49.00 ± 3.02	69.12 ± 1.83	68.94 ± 2.22

(7) PTFS [58] integrates discriminative partition information and applies self-paced learning strategies to improve unsupervised feature selection performance.
(8) SLNMF [54] utilizes a soft-label matrix based on local distance for supervision, and employs linear regression to correlate low-dimensional representations with label space, which effectively reduces redundancy, outliers, and noise.

4.2. Experimental setup

In this study, we assess the informativeness of the selected features through systematic clustering experiments. The evaluation follows a structured criterion: initially, features extracted from multi-view datasets are ranked using a variety of feature selection methods. Subsequently, the top k features are selected in descending order, with k varying over the set $\{10, 20, 30, \ldots, 280, 290, 300\}$, to construct a series of reduced datasets. Each dataset is then subjected to k-means clustering, which produces 20 independent clustering results per k. The predicted clusters are compared with ground truth labels, and the mean performance across these 20 runs is reported. Given the well-known sensitivity of k-means to initialization, this repetition enhances the robustness and reliability of our evaluation.

The primary aim of this study is to evaluate the effectiveness of the proposed method in addressing nonlinear problems, rather than focusing on the choice of kernel functions or parameter tuning. Accordingly, we adopt the widely used Gaussian kernel function, i.e., $K(x, y) =$

$\exp\left(-\frac{\|x-y\|^2}{td_{\max}^2}\right)$, where $d_{\max}$ denotes the maximum distance between samples, to capture the nonlinear structure of the data. To limit the number of parameters, a fixed-parameter strategy is applied with $t = 1$. In future work, we intend to further investigate the influence of alternative kernel functions and parameter settings on model performance.

To provide a comprehensive performance assessment, six widely accepted metrics are employed: accuracy (ACC), normalized mutual information (NMI), adjusted Rand index (ARI), F1 score, precision, and recall, with detailed descriptions available in [59]. For all metrics, higher values indicate the superior performance. Parameter settings for comparative feature selection algorithms are adopted from their respective original studies to ensure fairness. Specifically, for CCSFS, parameters β and γ are tuned across $\{2^3, 2^5, 2^7, 2^9, 2^{11}\}$, while λ is varied within $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$. For CDMvFS, β spans the same range $\{2^3, 2^5, 2^7, 2^9, 2^{11}\}$, and γ is explored over $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$. In our method, the parameter ρ is adjusted within $\{0.2, 0.4, 0.6, 0.8, 1\}$. To further guarantee impartiality, the remaining hyperparameters across all methods are varied over the range $\{10^{-2}, 10^{-1}, 1, 10^1, 10^2\}$. This setting is consistent with the parameter ranges used in baseline methods such as CCSFS and CDMvFS, which ensures the fairness of comparative experiments.

4.3. Experimental results

Table 4 provides a detailed summary of the experimental outcomes assessed using six standard evaluation metrics. The highest scores for each metric are marked in bold, while the second-best results are underlined for clarity. Figs. 3 and 4 further illustrate the impact of the varying number of selected features on clustering performance. Meanwhile, the results of the paired t-test are shown in Tables 5 and 6. Across the outdoor_scene_new, ORL, and handwritten datasets, the proposed method consistently achieves the superior performance compared to its counterparts. Notably, CDMvFS and PTFS also exhibit competitive results. In terms of ACC, our method achieves relative improvements of 3.19 %, 2.71 %, and 0.91 % over the second-best methods on the outdoor_scene_new, ORL, and handwritten datasets, respectively. With respect to NMI, performance gains of 0.53 %, 1.83 %, and 0.91 % are recorded. ARI is enhanced by 2.1 %, 3.07 %, and 3.69 %, while the F-score sees improvements of 1.7 %, 2.99 %, and 3.32 %. For precision, the proposed method outperforms the closest competitors by 2.41 %, 2.99 %, and 2.35 %, respectively.

From these tables and figures, we can draw the following conclusions.

(1) The paired t test results confirm GDFS's superior performance across most datasets. For ACC, GDFS showed significant improvement $h = 1$ in most cases(25/28), with particularly excellent results on Outdoor Scene, ORL, handwritten and 3Sources. A similar trend is seen in NMI, where GDFS achieves $h = 1$ in most comparisons(24/28), A paired t-test between GDFS and PTFS yielded a result of $h = 0$ on MSRCV1, Yale, and WebKB datasets, which indicates that the observed performance differences are not statistically significant at the 5 % level. Furthermore, GDFS remains highly competitive and demonstrates consistent advantages across other datasets.

(2) GDFS consistently ranks among the top performers across most experimental metrics, with particularly strong results on image datasets like ORL and Outdoor Scene. Its effectiveness stems from the joint modeling of local and global graph structures, nonlinear kernel mapping, and the integration of feature selection with consensus clustering. The incorporation of low-rank tensor constraints further enhances robustness by capturing cross-view consistency and reducing noise. GDFS also performs competitively on other image datasets like MSRCV1 and Yale, which confirms its generalizability. While PTFS leverages a statistics-based adaptive self-paced strategy, GDFS achieves comparable results without relying on external priors, underscoring its simplicity and effectiveness.

(3) GDFS achieves the best performance across all metrics on the Handwritten dataset, which highlights the effectiveness of its design. The dataset's clear class separation favors view-invariant and structurally consistent pseudo-label learning. By integrating graph structure modeling with consensus clustering, GDFS learns highly discriminative shared labels, and leads to strong results. While CDMvFS also performs well, GDFS further benefits from low-rank tensor constraints and kernel-based local structure modeling, which offer better robustness and more stable performance.

(4) On the WebKB dataset, GDFS achieves the best or second-best performance across four evaluation metrics. Notably, PTFS proves to be an effective method, attaining the highest scores in NMI and ARI. These results indicate that our approach is capable of effectively grouping samples into their correct categories. However, due to the blurred boundaries between classes and the presence of local noise in the WebKB dataset, some individual metrics may exhibit suboptimal performance.

4.4. Noise robustness

To evaluate the robustness of our method against additive Gaussian noise, we conducted experiments on the WebKB dataset by injecting scaled noise sampled from $\mathcal{N}(0, 1)$ with progressively increasing scaling factors (noise levels) of 0.1, 0.3, and 0.5. Here, a noise level of α indicates that the additive noise is $\alpha \cdot \mathcal{N}(0, 1)$, where $\alpha = 0$ corresponds to the original clean data. As shown in Table 7, the performance exhibits a graceful degradation with increasing noise intensity. Under low noise ($\alpha = 0.1$), the method is nearly identical to the clean case (ACC: 77.76 vs. 77.78), which demonstrates insensitivity to small perturbations. At moderate noise ($\alpha = 0.3$), the accuracy remains competitive at 76.28, with NMI and F-score declining by less than 4 %, respectively. Even under high noise ($\alpha = 0.5$), the method maintains an ACC of 72.98, with all metrics showing consistently low variance. These results suggest that our approach is robust to graded noise corruption, with performance degradation scaling predictably with noise intensity. This property is critical for real-world applications where data quality may vary.

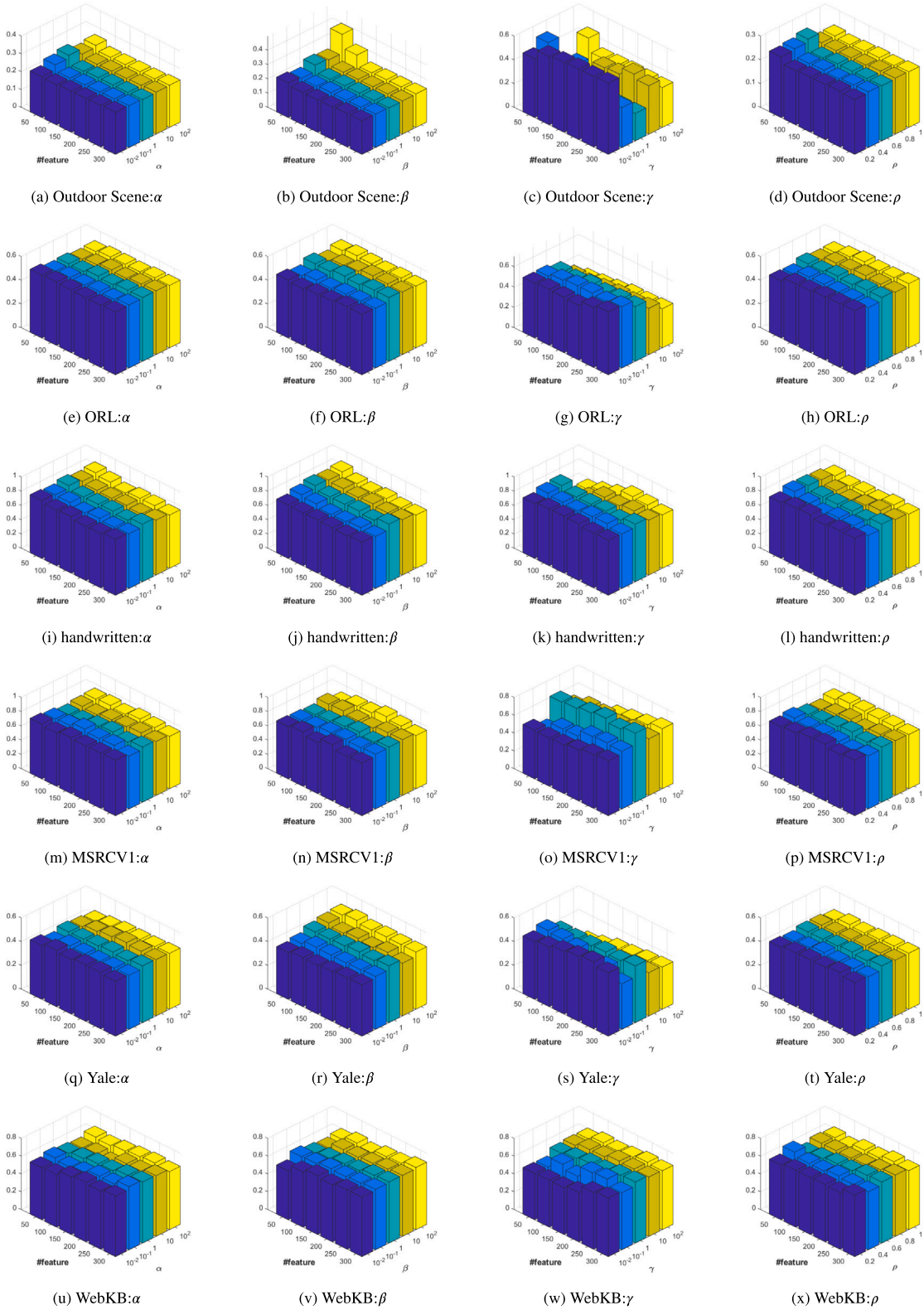
4.5. Parameter sensitivity

To assess the sensitivity of the proposed algorithm to its four manually configured parameters, we performed a series of controlled experiments, each aimed at evaluating the effect of a single parameter in isolation. In each experiment, one parameter was systematically varied while the remaining three were fixed at the midpoints of their respective predefined ranges. For example, when examining the impact of α, the other parameters were held constant, and the number of selected features was adjusted across the set $\{50, 100, 150, 200, 250, 300\}$ to explore performance across different feature dimensionalities. The corresponding results are visualized in Fig. 5.

The findings indicate that parameters α and ρ have a relatively minor influence on clustering performance across various datasets, which suggests that the algorithm demonstrates a degree of insensitivity to their specific settings. In contrast, β exhibits a more significant impact, with larger values generally leading to enhanced performance. The influence of γ appears to be dataset-specific. For instance, in image-based datasets such as outdoor_scene_new, ORL, and handwritten, lower values of γ tend to yield better results. However, for structured datasets like WebKB, higher values of γ are preferable. These observations underscore the importance of parameter sensitivity tuning in optimizing performance in multi-view learning applications.

4.6. Convergence study

Fig. 6 illustrates the convergence behavior of the proposed GDFS algorithm. Owing to the multi-block structure inherent in



(a) Outdoor Scene:α (b) Outdoor Scene:β (c) Outdoor Scene:γ (d) Outdoor Scene:ρ

(e) ORL:α (f) ORL:β (g) ORL:γ (h) ORL:ρ

(i) handwritten:α (j) handwritten:β (k) handwritten:γ (l) handwritten:ρ

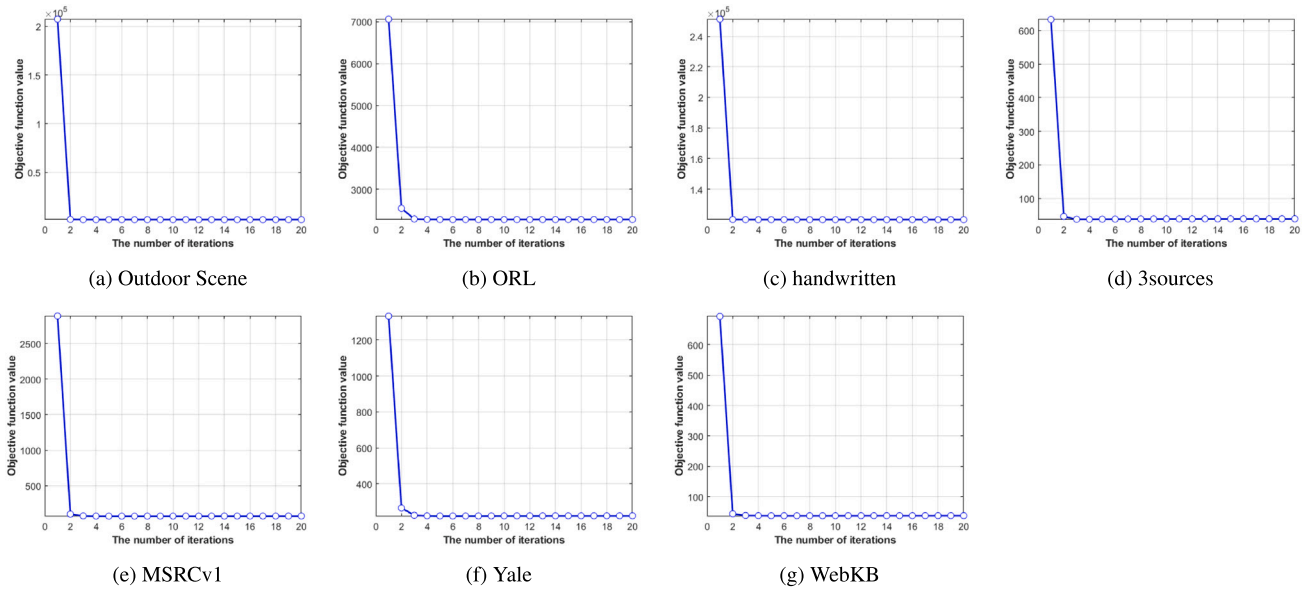
(m) MSRCV1:α (n) MSRCV1:β (o) MSRCV1:γ (p) MSRCV1:ρ

(q) Yale:α (r) Yale:β (s) Yale:γ (t) Yale:ρ

(u) WebKB:α (v) WebKB:β (w) WebKB:γ (x) WebKB:ρ

Fig. 5. Parameter sensitivity on different datasets.

Algorithm 1, which comprises five interdependent sub-problems, deriving a theoretical convergence guarantee remains a challenging task. Nevertheless, each sub-problem can be independently optimized to its respective minimum, which contributes to the algorithm's overall stability. Empirical evidence across diverse datasets confirms that GDFS converges consistently, with a rapid decline in the objective function observed within the first five iterations, followed by a steady convergence trend. These findings collectively affirm the practical convergence efficiency and reliability of the proposed method.



(a) Outdoor Scene (b) ORL (c) handwritten (d) 3sources

(e) MSRCv1 (f) Yale (g) WebKB

Fig. 6. Convergence study on different datasets.

5. Conclusion

In this work, we propose a novel multi-view unsupervised feature selection framework, termed GDFS, which effectively integrates local and global graph learning within a unified structure. By jointly modeling nonlinear local relationships in a kernel space and global structures from projected low-dimensional representations, GDFS captures both fine-grained and holistic data characteristics. Additionally, a graph discrepancy term and a low-rank tensor constraint are introduced to enhance inter-view consistency and suppress noise, while a consensus clustering matrix provides pseudo-label supervision for more robust feature selection. Although GDFS demonstrates strong performance across multiple benchmark datasets, it has several limitations: (1) it treats all views equally, which ignores their varying importance; (2) its computational complexity increases linearly with data size, which hinders scalability. In future work, we aim to address these issues by introducing an attention-based view-weighting strategy, and employing anchor graph techniques to reduce time complexity.

CRediT authorship contribution statement

Yiwan Xu: Writing – original draft, Visualization, Validation, Software. **Xijiong Xie:** Writing – review & editing, Supervision, Funding acquisition. **Xianliang Jiang:** Writing – review & editing, Supervision. **Yujie Xiong:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is supported by National Natural Science Foundation of China (No. 61906101). It is also supported by the Ningbo Municipal Natural Science Foundation of China (No. 2023J115).

Data availability

Data will be made available upon request.

References

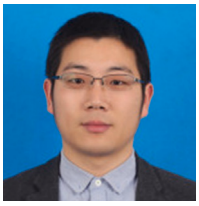
[1] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, 2005, pp. 886–893.
[3] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[4] J. Lv, Z. Kang, B. Wang, L. Ji, Z. Xu, Multi-view subspace clustering via partition fusion, Inf. Sci. 560 (2021) 410–423.
[5] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: a data perspective, ACM Comput. Surv. 50 (6) (2017) 1–45.
[6] Y. Guo, Z. Zhang, F. Tang, Feature selection with kernelized multi-class support vector machine, Pattern Recognit. 117 (2021) 107988.
[7] B. Tu, C. Zhou, J. Peng, G. Zhang, Y. Peng, Feature extraction via joint adaptive structure density for hyperspectral imagery classification, IEEE Trans. Instrum. Meas. 70 (2021) 1–16.
[8] H. Chen, X. Xie, D. Li, Semi-supervised learning with deep laplacian support vector machine, Pattern Anal. Appl. 28 (1) (2025) 1–13.
[9] R. Shang, Y. Meng, W. Wang, F. Shang, L. Jiao, Local discriminative based sparse subspace learning for feature selection, Pattern Recognit. 92 (2019) 219–230.
[10] J. Zheng, C. Luo, T. Li, H. Chen, A novel hierarchical feature selection method based on large margin nearest neighbor learning, Neurocomputing 497 (2022) 1–12.
[11] Z. Zhang, L. Shao, Y. Xu, L. Liu, J. Yang, Marginal representation learning with graph structure self-adaptation, IEEE Trans. Neural Netw. Learn. Syst. 29 (10) (2017) 4645–4659.
[12] L. Fu, Z. Li, Q. Ye, H. Yin, Q. Liu, X. Chen, X. Fan, W. Yang, G. Yang, Learning robust discriminant subspace based on joint l, -and l, -norm distance metrics, IEEE Trans. Neural Netw. Learn. Syst. 33 (1) (2020) 130–144.
[13] Y. Hu, J.-X. Liu, Y.-L. Gao, J. Shang, Dstpca: double-sparse constrained tensor principal component analysis method for feature selection, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (4) (2019) 1481–1491.
[14] Y. Huang, Z. Shen, Y. Cai, X. Yi, D. Wang, F. Lv, T. Li, C 2 imufs: complementary and consensus learning-based incomplete multi-view unsupervised feature selection, IEEE Trans. Knowl. Data Eng. 35 (10) (2023) 10681–10694.
[15] C. Tang, J. Chen, X. Liu, M. Li, P. Wang, M. Wang, P. Lu, Consensus learning guided multi-view unsupervised feature selection, Knowl.-Based Syst. 160 (2018) 49–60.
[16] Q. Lin, M. Men, L. Yang, P. Zhong, A supervised multi-view feature selection method based on locally sparse regularization and block computing, Inf. Sci. 582 (2022) 146–166.
[17] Q. Lin, L. Yang, P. Zhong, H. Zou, Robust supervised multi-view feature selection with weighted shared loss and maximum margin criterion, Knowl.-Based Syst. 229 (2021) 107331.
[18] C. Shi, Z. Gu, C. Duan, Q. Tian, Multi-view adaptive semi-supervised feature selection with the self-paced learning, Signal Process. 168 (2020) 107332.
[19] B. Jiang, X. Wu, X. Zhou, Y. Liu, A.G. Cohn, W. Sheng, H. Chen, Semi-supervised multiview feature selection with adaptive graph learning, IEEE Trans. Neural Netw. Learn. Syst. 35 (3) (2024) 3615–3629.
[20] B. Jiang, J. Liu, Z. Wang, C. Zhang, J. Yang, Y. Wang, W. Sheng, W. Ding, Semi-supervised multi-view feature selection with adaptive similarity fusion and learning, Pattern Recognit. 159 (2025) 111159.

[21] H. Liu, H. Mao, Y. Fu, Robust multi-view feature selection, in: 2016 IEEE 16th International Conference on Data Mining (ICDM), 2016, pp. 281–290.
[22] H. Zhang, D. Wu, F. Nie, R. Wang, X. Li, Multilevel projections with adaptive neighbor graph for unsupervised multi-view feature selection, Inf. Fusion 70 (2021) 129–140.
[23] S.-G. Fang, D. Huang, C.-D. Wang, Y. Tang, Joint multi-view unsupervised feature selection and graph learning, IEEE Trans. Emerg. Top. Comput. Intell. 8 (1) (2023) 16–31.
[24] C. Zhang, Y. Fang, X. Liang, H. Zhang, P. Zhou, X. Wu, J. Yang, B. Jiang, W. Sheng, Efficient multi view unsupervised feature selection with adaptive structure learning and inference, in: Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, 2024.
[25] S. Zhou, P. Song, Consistency–exclusivity guided unsupervised multi-view feature selection, Neurocomputing 569 (2024) 127119.
[26] Y. Huang, M. Lu, W. Huang, X. Yi, T. Li, Time-fs: joint learning of tensorial incomplete multi-view unsupervised feature selection and missing-view imputation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, 2025, pp. 17503–17510.
[27] X. Yang, H. Che, M.-F. Leung, S. Wen, Unbalanced incomplete multiview unsupervised feature selection with low-redundancy constraint in low-dimensional space, IEEE Trans. Ind. Inf. (2024).
[28] Z. Cao, X. Xie, Y. Li, Multi-view unsupervised feature selection with consensus partition and diverse graph, Inf. Sci. 661 (2024) 120178.
[29] F. Nie, W. Zhu, X. Li, Unsupervised feature selection with structured graph optimization, in: Proceedings of the AAAI conference on artificial intelligence, vol. 30, 2016.
[30] J. Liu, X. Liu, Y. Yang, Q. Liao, Y. Xia, Contrastive multi-view kernel learning, IEEE Trans. Pattern Anal. Mach. Intell. 45 (8) (2023) 9552–9566.
[31] H.Q. Minh, P. Niyogi, Y. Yao, Mercer's theorem, feature maps, and smoothing, in: International Conference on Computational Learning Theory, Springer, 2006, pp. 154–168.
[32] S. Sonnenburg, G. Rätsch, C. Schäfer, B. Schölkopf, Large scale multiple kernel learning, J. Mach. Learn. Res. 7 (2006) 1531–1565.
[33] H.-C. Huang, Y.-Y. Chuang, C.-S. Chen, Multiple kernel fuzzy clustering, IEEE Trans. Fuzzy Syst. 20 (1) (2011) 120–134.
[34] M. Kloft, U. Rückert, P.L. Bartlett, A unifying view of multiple kernel learning, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part II 21, Springer, 2010, pp. 66–81.
[35] X. Liu, Y. Dou, J. Yin, L. Wang, E. Zhu, Multiple kernel k-means clustering with matrix-induced regularization, in: Proceedings of the AAAI conference on artificial intelligence, vol. 30, 2016.
[36] J. Liu, X. Liu, J. Xiong, Q. Liao, S. Zhou, S. Wang, Y. Yang, Optimal neighborhood multiple kernel clustering with adaptive local kernels, IEEE Trans. Knowl. Data Eng. 34 (6) (2020) 2872–2885.
[37] M. Bouafia, D. Benterki, A. Yassine, An efficient parameterized logarithmic kernel function for linear optimization, Optim. Lett. 12 (2018) 1079–1097.
[38] J. Miao, T. Yang, L. Sun, X. Fei, L. Niu, Y. Shi, Graph regularized locally linear embedding for unsupervised feature selection, Pattern Recognit. 122 (2022) 108299.
[39] J.-S. Wu, M.-X. Song, W. Min, J.-H. Lai, W.-S. Zheng, Joint adaptive manifold and embedding learning for unsupervised feature selection, Pattern Recognit. 112 (2021) 107742.
[40] H. Yuan, J. Li, Y. Liang, Y.Y. Tang, Multi-view unsupervised feature selection with tensor low-rank minimization, Neurocomputing 487 (2022) 75–85.
[41] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint 2, 1-norms minimization, Adv. Neural Inf. Process. Syst. 23 (2010).
[42] T. Zhang, Y. Yuan, S.F. Liu, Two-step affinity matrix learning for multi-view subspace clustering, Expert Syst. Appl. 242 (May) (2024) 122765.1–122765.14.
[43] Y. Xie, D. Tao, W. Zhang, Y. Liu, L. Zhang, Y. Qu, On unifying multi-view self-representations for clustering by tensor multi-rank minimization, Int. J. Comput. Vis. 126 (2018) 1157–1179.
[44] J. Wu, Z. Lin, H. Zha, Essential tensor learning for multi-view spectral clustering, IEEE Trans. Image Process. 28 (12) (2019) 5910–5922.
[45] G.-Y. Zhang, Y.-R. Zhou, C.-D. Wang, D. Huang, X.-Y. He, Joint representation learning for multi-view subspace clustering, Expert Syst. Appl. 166 (2021) 113913.
[46] X. Yang, H. Che, M.-F. Leung, Tensor-based unsupervised feature selection for error-robust handling of unbalanced incomplete multi-view data, Inf. Fusion 114 (2025) 102693.
[47] Z. Kang, C. Peng, Q. Cheng, Clustering with adaptive manifold structure learning, in: 2017 IEEE 33rd International Conference on Data Engineering (ICDE), IEEE, 2017, pp. 79–82.
[48] Z. Cao, X. Xie, Structure learning with consensus label information for multi-view unsupervised feature selection, Expert Syst. Appl. 238 (2024) 121893.
[49] X. Xie, Z. Cao, F. Sun, Joint learning of graph and latent representation for unsupervised feature selection, Appl. Intell. 53 (21) (2023) 25282–25295.
[50] W. Hu, D. Tao, W. Zhang, Y. Xie, Y. Yang, The twist tensor nuclear norm for video completion, IEEE Trans. Neural Netw. Learn. Syst. 28 (12) (2017) 2961–2973.
[51] F. Nie, X. Wang, M. Jordan, H. Huang, The constrained laplacian rank algorithm for graph-based clustering, in: Proceedings of the AAAI conference on artificial intelligence, vol. 30, 2016.
[52] Z. Lin, M. Chen, Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, arXiv preprint arXiv:1009.5055, (2010).
[53] Y. Zhang, Recent advances in alternating direction methods: practice and theory, in: IPAM workshop on continuous optimization, vol. 385, 2010.
[54] S. Zhou, P. Song, Z. Song, L. Ji, Soft-label guided non-negative matrix factorization for unsupervised feature selection, Expert Syst. Appl. 216 (2023) 119468.
[55] C. Hou, F. Nie, H. Tao, D. Yi, Multi-view unsupervised feature selection with adaptive similarity and view weight, IEEE Trans. Knowl. Data Eng. 29 (9) (2017) 1998–2011.
[56] C. Tang, X. Zhu, X. Liu, L. Wang, Cross-view local structure preserved diversity and consensus learning for multi-view unsupervised feature selection, in: Proceedings of the AAAI Conference on artificial intelligence, vol. 33, 2019, pp. 5101–5108.
[57] Z. Cao, X. Xie, F. Sun, J. Qian, Consensus cluster structure guided multi-view unsupervised feature selection, Knowl.-Based Syst. 271 (2023) 110578.
[58] Z. Cao, X. Xie, Partition-level tensor learning-based multiview unsupervised feature selection, IEEE Trans. Neural Netw. Learn. Syst. 36 (7) (2025) 12799–12811.
[59] Y. Jia, H. Liu, J. Hou, S. Kwong, Q. Zhang, Multi-view spectral clustering tailored tensor low-rank representation, IEEE Trans. Circuits Syst. Video Technol. 31 (12) (2021) 4784–4797.

Author biography

Yiwan Xu is currently pursuing a master's degree at Ningbo University in Ningbo, China. Her research interests include kernel methods, feature selection and multi-view learning.

Xijiong Xie received his Ph.D. degree from the Department of Computer Science and Technology, East China Normal University in 2016. He is currently an Associate Professor at the School of Information Science and Engineering, Ningbo University, China. Dr. Xie was listed in the Top 2% Scientists by Stanford University and recognized as a leading talent in Ningbo City. He has published over 50 papers in renowned journals and conferences such as IEEE TKDE, IEEE TC, IEEE TNNLS, Information Fusion, KBS, PR, Information Sciences, ESWA, Neurocomputing and Applied Intelligence. His research interests include multi-view learning, clustering, dimensionality reduction, support vector machines, computer vision and deep learning.



Xianliang Jiang received the Ph.D. degree in Computer Science and Technology from the Zhejiang University in 2016. Before that he received the B.E. degree from the University of Science and Technology of China in 2009, and the M.S. degree from the Ningbo University in 2012. He is currently a lecturer and his research interests include network protocol design, video streaming algorithm, and intelligent Internet of Things.

Yujie Xiong received his Ph.D. in Computer Science from the East China Normal University in June 2018. He is currently an Associate Professor in the School of Electronic and Electrical Engineering at Shanghai University of Engineering Science, China. His research interests include biometric, document image analysis, and knowledge graph based ap plication, where he has published more than 30 academic publications.