Article:
Jiang, B, Wu, X, Zhou, X et al. (4 more authors) (2022) Semi-Supervised Multi-View Feature Selection with Adaptive Graph Learning. IEEE Transactions on Neural Networks and Learning Systems. ISSN 2162-237X

https://doi.org/10.1109/TNNLS.2022.3194957

Semi-Supervised Multi-View Feature Selection with Adaptive Graph Learning

Bingbing Jiang, Xingyu Wu, Xiren Zhou, Yi Liu, Anthony G. Cohn, Weiguo Sheng, *Member, IEEE*, and Huanhuan Chen, *Senior Member, IEEE*

***Abstract*—As data sources become ever more numerous with increased feature dimensionality, feature selection for multi-view data has become an important technique in machine learning. Semi-supervised multi-view feature selection focuses on the problem of how to obtain a discriminative feature subset from heterogeneous feature spaces in the case of abundant unlabeled data with little labeled data. Most existing methods suffer from unreliable similarity graph structure across different views since they separate the graph construction from feature selection and use the fixed graphs that are susceptible to noisy features. Furthermore, they directly concatenate multiple feature projections for feature selection, neglecting the contribution diversity among projections. To alleviate these problems, we present a semi-supervised multi-view feature selection (SMFS) to simultaneously select informative features and learn a unified graph through the data fusion from aspects of feature projection and similarity graph. Specifically, SMFS adaptively weights different feature projections and flexibly fuses them to form a joint weighted projection, preserving the complementarity and consensus of the original views. Moreover, an implicit graph fusion is devised to dynamically learn a compatible graph across views according to the similarity structure in the learned projection subspace, where the undesirable effects of noisy features are largely alleviated. A convergent method is derived to iteratively optimize SMFS. Experiments on various datasets validate the effectiveness and superiority of SMFS over state-of-the-art methods.**

***Index Terms*—Multi-view feature selection, semi-supervised learning, graph learning, multi-view data fusion.**

I. Introduction

In many real-world applications, such as scene classification, handwritten recognition and object detection, data with multiple representations from different sources are collected to form multi-view data. In multi-view data, each view naturally corresponds to a feature representation that contains independent statistical properties [1]–[4]. For instance, an image can be depicted by diverse types of features, such as HOG [5], LBP [6] GIST [7], etc. Although these multi-view features characterize data from different perspectives, their high dimensionality demands high computation and massive storage during data processing. Moreover, high-dimensional data inevitably contain some redundant or even irrelevant (noisy) features that might degrade the effectiveness of subsequent tasks [8], becoming the main challenge in data mining. As a topic of interest in multi-view learning, multi-view feature selection aims to obtain a lower-dimensional feature representation (i.e., feature subset) by removing the irrelevant and redundant features from the heterogeneous feature space. Due to the comprehensive representation and better interpretability for multi-view data, multi-view feature selection has caught the attention of many researchers [9]–[14].

The key challenge in multi-view feature selection is to effectively mine and exploit the consensus and complementarity among views to select the features that cover the original feature space well. Current multi-view feature selection techniques can be categorized into supervised, unsupervised and semi-supervised groups according to the availability of label information in data [15]. Supervised methods usually require sufficient labeled data to maintain promising performance. Unsupervised methods select features guided by the intrinsic structure of data, but often fail to identify some discriminative features due to the absence of label information. Abundant unlabeled data in the real world are expensive to label, therefore, it is desirable to develop multi-view semi-supervised feature selection that simultaneously exploits labeled and unlabeled data. Although semi-supervised multi-view learning has attracted much attention in recent years [16]–[20], to the best of our knowledge, relatively few efforts have been made on semi-supervised multi-view feature selection.

To identify relevant features from multi-view data where only a small proportion has been labeled, existing methods in the literature follow two different ways. The traditional way is to indiscriminately concatenate multiple features first and import the concatenated features into single-view models that mostly employ data distribution structure or a sparsity constraint to evaluate the importance of features [21]–[25]. For example, Zhao *et al* proposed to select features according to the label information and the local structure information of data [21]. Ma *et al.* proposed a structural feature selection with sparsity [24] which selects features by jointly exploiting an l_{21}-norm regularization and manifold learning. Chen *et al.* [25] proposed a semi-supervised feature selection via sparse rescaled linear square regression model, using a general l_{2p}-norm ($p \in (0, 1]$) to ensure the sparsity of the feature space. However, these methods are originally designed for single-view data such that they treat view-specific features equally and neglect the complementarity among views, weakening

This work was supported in part by the National Natural Science Foundation of China under Grants 62006065 and 61873082; and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY22F030004. Weiguo Sheng and Huanhuan Chen are both corresponding authors.

Bingbing Jiang, Yi Liu and Weiguo Sheng are with Hangzhou Normal University, Hangzhou 311121, China (e-mail: jiangbb@hznu.edu.cn;liuyi@hznu.edu.cn;w.sheng@ieee.org).

Xingyu Wu, Xiren Zhou, Huanhuan Chen are with University of Science and Technology of China, Hefei 230027, China (e-mail: xingyuwu@mail.ustc.edu.cn; zhou0612@ustc.edu.cn;hchen@ustc.edu.cn).

Anthony G. Cohn is with School of Computing, University of Leeds, Leeds, UK (e-mail: A.G.Cohn@leeds.ac.uk).

their effectiveness and applicability for multi-view scenarios.

To alleviate this issue, another way is to directly perform multi-view feature selection, thereby considering the diversity of different views. Representative methods include manifold regularized multi-view feature selection (MRMVFS) [10], multi-view Laplacian sparse feature selection (MLSFS) [26] and multi-view Hessian semi-supervised sparse feature selection [27]. These methods commonly follow two steps, i.e., construction of an individual graph for each view to describe the view-specific similarity structure and feature selection based on a weighted combination of multiple view-specific structures. Despite achieving better performance than single-view methods, these multi-view methods suffer from at least two deficiencies including: **a)** the view-specific similarity graph is simply derived from the original feature space where lots of redundant and noisy features inevitably exist. Moreover, the local similarity structure remains constant during the process of feature selection, making the learned graph suboptimal; **b)** these methods treat the feature projections of different views indiscriminately and directly use the projection matrix of concatenated features to select features, ignoring the relationship (i.e., the complementarity and consensus) between different views in the aspect of feature projection [28], [29]. Recently, Shi *et al.* proposed a multi-view adaptive semi-supervised feature selection (MASFS), which dynamically updates view-specific graphs according to the Euclidean distance between original data points [30]. However, the Euclidean distance is sensitive to the redundant and noisy features that usually exist in multi-view data leading to the unreliable graph [31]–[33], thereby affecting the effectiveness of selected features.

Motivated by the aforementioned problems, this paper proposes a novel multi-view feature selection algorithm, called *S*emi-supervised *M*ulti-view *F*eature *S*election with adaptive graph learning (SMFS). The main contributions of this paper are summarized as follows:

- We design a novel semi-supervised multi-view feature selection algorithm that comprehensively exploits multi-view data from the aspects of both the feature projections and the similarity graphs to adaptively learn a unified graph and simultaneously select discriminative features.
- We distinguish the feature projections of different views and adaptively coalesce them to form a joint weighted feature projection by merging the learned view weights into the view-specific projections, so that the complementarity and consistency among multiple views in the level of feature projection can be naturally preserved.
- We learn a reliable similarity graph across multiple views by virtue of the implicit graph fusion across views and the similarity structure in the projected feature subspace, which positively facilitates the feature selection.
- An effective and provably convergent optimization method is developed to solve the formulated objective function, and extensive experiments on various datasets are conducted to demonstrate the effectiveness of SMFS and its superiority over other state-of-the-art competitors.

The remainder of this paper is organized as follows. We first introduce the formulation of SMFS, including adaptive-weighting multi-view feature selection, adaptive graph learning and optimization procedure in Section II. Then Section III analyzes SMFS in three aspects. In Section IV, we conduct extensive experiments to validate the proposed SMFS algorithm on various multi-view datasets. Finally, we conclude this paper and propose some future directions in Section V.

II. Semi-Supervised Multi-View Feature Selection with Adaptive Graph Learning

In this section, we introduce the proposed SMFS algorithm. Firstly, the adaptive-weighting multi-view feature selection model is formulated in section II-A. Adaptive graph fusion and learning is developed in section II-B. Section II-C presents the unified framework integrating the multi-view feature selection and adaptive graph fusion and learning. Finally, Section II-D details the optimization procedures of SMFS.

A. Adaptive-Weighting Multi-View Feature Selection Model

Recently, a multi-view fusion model has been widely used with a unified formulation as follows:

$$\min_{\pi_v \geq 0, \sum_v \pi_v = 1} \sum_{v=1}^{V} \pi_v^{\eta} f_v(x), \tag{1}$$

where π_v is the weight of the v-th view, $f_v(\cdot)$ is a problem-specific function, x is a task-dependent variable, V is the number of views, and hyper-parameter $\eta > 1$ controls the distribution of $\{\pi_v\}_{v=1}^{V}$. Multi-view learning can be implemented in various ways using different functions and variables.

In the machine learning field, a linear least-squares regression model is frequently utilized to learn a projection subspace according to the prediction label $\mathbf{F} \in \mathbb{R}^{n \times c}$ (n and c are the numbers of samples and classes, respectively), thereby preserving the discriminative information of training data [34]. Accordingly, $f_v(\cdot)$ can be materialized as a loss function which encodes the mismatch between the linear projection $\mathbf{X}_v^T\mathbf{W}_v + \mathbf{1}\mathbf{b}_v^T$ and $\mathbf{F}$, where $\mathbf{b}_v \in \mathbb{R}^{c \times 1}$ denotes the view-specific bias vector and $\mathbf{W}_v \in \mathbb{R}^{d_v \times c}$ (d_v is the dimensionality of the v-th view) is the feature projection matrix that maps the original features $\mathbf{X}_v \in \mathbb{R}^{d_v \times n}$ into the subspace. For simplicity, $\mathbf{b}_v$ is absorbed into $\mathbf{W}_v$ by adding $\mathbf{1} \in \mathbb{R}^{n \times 1}$ as an additional row of $\mathbf{X}_v$ to reform Eq. (1) as below:

$$\min_{\mathbf{F}, \mathbf{W}_v, \pi_v \geq 0, \sum_v \pi_v = 1} \sum_{v=1}^{V} \pi_v^{\eta} \|\mathbf{X}_v^T\mathbf{W}_v - \mathbf{F}\|_F^2. \tag{2}$$

By keeping $\mathbf{W}_v$ and π_v fixed, we take the derivative of Eq. (2) with respect to (w.r.t.) $\mathbf{F}$ and set it to be zero:

$$\sum_{v=1}^{V} \pi_v^{\eta}\left(\mathbf{F} - \mathbf{X}_v^T\mathbf{W}_v\right) = 0 \Longrightarrow \begin{cases} \mathbf{F} = \sum\limits_{v=1}^{V} \alpha_v \mathbf{X}_v^T\mathbf{W}_v \\ \alpha_v = \pi_v^{\eta} / \sum\limits_{v=1}^{V} \pi_v^{\eta} \end{cases}, \tag{3}$$

where α_v can be viewed as the non-negative weight for the v-th projection subspace $\mathbf{X}_v^T\mathbf{W}_v$ and $\sum_{v=1}^{V} \alpha_v = 1$. The solution to $\mathbf{F}$ is achieved by merging the single-view projection subspace set $\{\mathbf{X}_v^T\mathbf{W}_v\}_{v=1}^{V}$ in an optimal combination. Note that $\{\alpha_v\}_{v=1}^{V}$ are linearly imposed on the single-view projections, which might lead to the trivial solution, that is $\alpha_v = 1$ only

for the best one and $\alpha_v = 0$ for others, thereby impeding the possibility to take full advantage of the consistent and complementary information contained in multiple views.

To better cope with this problem, the equality constraint of $\mathbf{F}$ is relaxed to a quadratic penalty term by introducing a flexible regression residue (i.e., $\mathbf{F} - \sum_{v=1}^{V} \alpha_v \mathbf{X}_v^T \mathbf{W}_v$) to model the mismatch between $\mathbf{F}$ and the weighted combination projection $\sum_{v=1}^{V} \alpha_v \mathbf{X}_v^T \mathbf{W}_v$. Meanwhile, considering that each row of $\mathbf{W}_v$ reflects the importance of the corresponding feature, if $\mathbf{W}_v$ is sparse in rows, feature selection can be naturally performed by selecting the features associated with the non-zero rows in $\mathbf{W}_v$ [35]. Owing to the efficacy shown in the state-of-the-art works [11], [36], the $l_{2,1}$-norm constraint is imposed on $\mathbf{W}_v$ to ensure that $\mathbf{W}_v$ is row-sparse. Therefore, a feature selection model that compatibly fuses multiple views from the level of projection subspace is put forward:

$$\min_{\mathbf{F},\mathbf{W}_v,\boldsymbol{\alpha}\geq 0,\mathbf{1}^T\boldsymbol{\alpha}=1} \|\mathbf{F} - \sum_{v=1}^{V} \alpha_v \mathbf{X}_v^T \mathbf{W}_v\|_F^2 + \lambda \sum_{v=1}^{V} \|\mathbf{W}_v\|_{2,1}, \quad (4)$$

where $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_V] \in \mathbb{R}^V$ is a weight vector. Different from most existing studies, the proposed model learns view-specific feature projection subspaces $\{\mathbf{X}_v^T \mathbf{W}_v\}_{v=1}^V$ for each view and discriminates different projections with the view weights $\{\alpha_v\}_{v=1}^V$, which takes full advantage of the complementary information from different projections. To facilitate feature projections consensus, the prediction label matrix $\mathbf{F}$ is employed as the common regression target across views, which maximizes the consistency among view-specific subspaces. Moreover, considering that the weight α_v depends on the discrepancy between $\mathbf{F}$ and $\mathbf{X}_v^T \mathbf{W}_v$ such that none of single-view feature projections can outperform others. By modeling this discrepancy, α_v can be determined adaptively. Therefore, this fusion manner not only avoids the trivial solution of $\boldsymbol{\alpha}$ but also releases the model from the extra hyper-parameter η.

However, the objective function in Eq. (4) is difficult to be optimized due to the introduction of $\boldsymbol{\alpha}$. To solve this problem, we merge α_v into its associated feature projection matrix $\mathbf{W}_v = [\mathbf{w}_v^1, \mathbf{w}_v^2, \cdots, \mathbf{w}_v^{d_v}]$ as $\alpha_v \mathbf{W}_v = \widetilde{\mathbf{W}}_v$, where $\mathbf{w}_v^i$ is the i-th row of $\mathbf{W}_v$ and $\widetilde{\mathbf{W}}_v = [\alpha_v \mathbf{w}_v^1, \alpha_v \mathbf{w}_v^2, \cdots, \alpha_v \mathbf{w}_v^{d_v}] \in \mathbb{R}^{d_v \times c}$ denotes the weighted feature projection matrix. According to the definition of $l_{2,1}$-norm, we have:

$$\begin{aligned} \|\widetilde{\mathbf{W}}_v\|_{2,1} &= \sum_{i=1}^{dv} \sqrt{\alpha_v \mathbf{w}_v^i (\alpha_v \mathbf{w}_v^i)^T} = \alpha_v \sum_{i=1}^{dv} \sqrt{\mathbf{w}_v^i (\mathbf{w}_v^i)^T} = \alpha_v \|\mathbf{W}_v\|_{2,1} \\ \Longrightarrow \sum_{v=1}^{V} \|\mathbf{W}_v\|_{2,1} &= \sum_{v=1}^{V} \frac{\|\widetilde{\mathbf{W}}_v\|_{2,1}}{\alpha_v}. \end{aligned} \quad (5)$$

Therefore, the proposed multi-view feature selection model in Eq. (4) can be transformed into:

$$\min_{\mathbf{F},\widetilde{\mathbf{W}},\boldsymbol{\alpha}\geq 0,\mathbf{1}^T\boldsymbol{\alpha}=1} \|\mathbf{F} - \mathbf{X}^T \widetilde{\mathbf{W}}\|_F^2 + \lambda \sum_{v=1}^{V} \frac{\|\widetilde{\mathbf{W}}_v\|_{2,1}}{\alpha_v}, \quad (6)$$

where $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_V]^T \in \mathbb{R}^{d \times n}$ is the concatenated feature matrix, $\widetilde{\mathbf{W}} = [\widetilde{\mathbf{W}}_1, \cdots, \widetilde{\mathbf{W}}_V]^T \in \mathbb{R}^{d \times c}$ denotes a joint weighted feature projection matrix across all V views, and $d = \sum_{v=1}^{V} d_v$ is the dimension of features across all V views. Different from previous methods [10], [26], [30], the proposed multi-view feature selection model not only distinguishes the feature projections derived from different views, but also coalesces them in an adaptive-weighting way, thereby providing a comprehensive representation and feature projection-level fusion compatible across all views. Moreover, we can efficiently optimize Eq. (6) by merging α_v into the corresponding projection matrix, such that the correlation and complementarity among multiple views can be naturally preserved in the joint weighted projection subspace $\mathbf{X}^T \widetilde{\mathbf{W}}$.

B. Adaptive Graph Fusion and Learning

In real-world applications, the unavailability of labeled data motivates feature selection to evaluate features based on the similarity structure of data. Consequently, learning a graph that effectively preserves the structure is critical for selecting informative features [36]. To avoid the separation of graph construction and feature selection, we plan to perform feature selection and graph learning simultaneously, such that the sample similarity in the projected feature subspace is also taken into account, alleviating the adverse impact of redundant and irrelevant features (in the original feature space). Toward this end, two important factors are considered for adaptive graph fusion and learning: **a)** the graph fusion should explore the intrinsic structure across multiple views, so that the fused graph not only mines the similarity structure within each view, but also captures the structure compatible across multiple views; **b)** the similarity information directly derived from the original feature space might not be fully reliable, making the learned feature selection matrix suboptimal and thereby degrading the performance. To achieve them, an adaptive graph fusion and learning model is proposed as follows:

$$\min_{\mathbf{S}\mathbf{1}=\mathbf{1},\mathbf{S}\geq 0} \sum_{v=1}^{V} \|\mathbf{S} - \mathbf{S}^v\|_F + \mu \sum_{i,j=1}^{n} s_{ij} \|\widetilde{\mathbf{W}}^T \mathbf{x}_i - \widetilde{\mathbf{W}}^T \mathbf{x}_j\|_2^2, \quad (7)$$

where $\{\mathbf{S}^v\}_{v=1}^V$ denotes the view-specific graphs constructed from the original feature space [37], s_{ij} is the similarity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\mu > 0$ is the regularization parameter. The first term implicitly fuses previously built single-view graphs $\{\mathbf{S}^v\}_{v=1}^V$ to learn a unified graph $\mathbf{S}$ through automatically assigning weights according to the matching degrees between $\mathbf{S}$ and $\{\mathbf{S}^v\}_{v=1}^V$. The second term can be viewed as a weighted ℓ_1-norm constraint on $\mathbf{S}$, which enables Eq. (7) to adaptively select several projected neighbors. By virtue of the weighted concatenated feature projection matrix $\widetilde{\mathbf{W}}$, the graph $\mathbf{S}$ adaptively fused from $\{\mathbf{S}^v\}_{v=1}^V$ is also guided by the sample similarity in projected feature subspace. In this way, multi-view feature selection and similarity graph learning can benefit from each other in a mutual reinforcement manner.

C. SMFS Algorithm

We propose a new semi-supervised multi-view feature selection (SMFS) algorithm by combining adaptive graph learning and multi-view feature selection based on the preliminary formulations described in Sections II-A and II-B. It is pertinent to note that the feature selection matrix $\widetilde{\mathbf{W}}$ obtained from Eq. (6) is affected by the prediction label matrix $\mathbf{F}$. To select informative and discriminative features, $\mathbf{F}$ should vary

smoothly along with the learned graph $\mathbf{S}$ and be consistent with the known labels of l labeled samples (i.e., $\mathbf{Y}_l$) according to the semi-supervised setting [38]–[40]. Specifically, $\mathbf{F}$ can be further constrained by graph-based label propagation [41]:

$$\min_{\mathbf{F}} \ \mathrm{Tr}(\mathbf{F}^T\mathbf{L_S}\mathbf{F}) + \mathrm{Tr}\big((\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})\big), \tag{8}$$

where $\mathbf{L_S} \in \mathbb{R}^{n\times n}$ is the Laplacian matrix of $\mathbf{S}$, $\mathbf{Y} = [\mathbf{Y}_l; \mathbf{0}]^T \in \mathbb{R}^{n\times c}$, and $\mathbf{U} \in \mathbb{R}^{n\times n}$ is a diagonal matrix in which diagonal element U_{ii} is a very large number if $\mathbf{x}_i$ is a labeled sample and 1 otherwise. $\mathbf{L_S} = \mathbf{D_S} - (\mathbf{S}^T + \mathbf{S})/2$, where $\mathbf{D_S}$ is the diagonal degree matrix with the i-th diagonal element $\sum_{j=1}^{n}(s_{ij} + s_{ji})/2$. In addition to constraining the smoothness and the label consistency of $\mathbf{F}$, Eq. (8) ties the graph learning and the feature selection processes together. Based on the multi-view feature selection and the graph learning proposed in Eq. (6) and Eq. (7), respectively, we formulate the final objective function of SMFS as follows:

$$\begin{aligned}
&\min_{\mathbf{F},\widetilde{\mathbf{W}},\mathbf{S},\boldsymbol{\alpha}} \ \mathrm{Tr}(\mathbf{F}^T\mathbf{L_S}\mathbf{F}) + \mathrm{Tr}\big((\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})\big) \\
&+ \gamma\Big(\|\mathbf{F}-\mathbf{X}^T\widetilde{\mathbf{W}}\|_F^2 + \lambda\sum_{v=1}^{V}\frac{\|\widetilde{\mathbf{W}}_v\|_{2,1}}{\alpha_v}\Big) \\
&+ \beta\Big(\sum_{v=1}^{V}\|\mathbf{S}-\mathbf{S}^v\|_F + \mu\sum_{i,j=1}^{n} s_{ij}\|\widetilde{\mathbf{W}}^T\mathbf{x}_i - \widetilde{\mathbf{W}}^T\mathbf{x}_j\|_2^2\Big) \\
&\text{s.t. } \ \boldsymbol{\alpha} \geq 0, \boldsymbol{\alpha}^T\mathbf{1} = 1, \ \mathbf{S1} = \mathbf{1}, \mathbf{S} \geq 0,
\end{aligned} \tag{9}$$

where γ and β are the parameters to balance projection learning and graph fusion, respectively. Different from existing methods, SMFS not only discriminates multiple feature projections by virtue of the adaptive weight $\boldsymbol{\alpha}$, but also preserves the consensus and complementarity in the projected subspace by merging $\boldsymbol{\alpha}$ into the joint projection matrix $\widetilde{\mathbf{W}}$. Furthermore, by projecting the original data $\mathbf{X}$ into the subspace, SMFS performs feature selection and graph learning based on the sample similarity in the projected subspace, thereby alleviating the undesirable effect of irrelevant and redundant features.

D. Optimization Procedure

It is important to emphasize that Eq. (9) is difficult to solve immediately since it contains the unsmooth $l_{2,1}$-norm and is not a jointly convex problem with respect to all variables. To obtain the optimal solution, we design an iterative optimization method which works in iterations to alternately optimize a variable by fixing the values of other variables and guarantees the monotonic decrease of the objective function in Eq. (9). This can be achieved via the update rules given below:

• **Update F:** when $\widetilde{\mathbf{W}}$, $\boldsymbol{\alpha}$ and $\mathbf{S}$ are fixed, the subproblem of Eq. (9) is simplified into:

$$\min_{\mathbf{F}} \mathrm{Tr}(\mathbf{F}^T\mathbf{L_S}\mathbf{F}) + \mathrm{Tr}\big((\mathbf{F}-\mathbf{Y})^T\mathbf{U}(\mathbf{F}-\mathbf{Y})\big) + \gamma\|\mathbf{F}-\mathbf{X}^T\widetilde{\mathbf{W}}\|_F^2. \tag{10}$$

Taking the partial derivative of Eq. (10) w.r.t. $\mathbf{F}$ and setting it to be zero, we have:

$$\gamma(\mathbf{F}-\mathbf{X}^T\widetilde{\mathbf{W}}) + \mathbf{L_S}\mathbf{F} + \mathbf{U}(\mathbf{F}-\mathbf{Y}) = 0 \Longrightarrow \mathbf{F} = \mathbf{PQ}, \tag{11}$$

where $\mathbf{P} = (\gamma\mathbf{I} + \mathbf{L_S} + \mathbf{U})^{-1}$ and $\mathbf{Q} = \mathbf{UY} + \gamma\mathbf{X}^T\widetilde{\mathbf{W}}$.

• **Update $\widetilde{\mathbf{W}}$:** replacing $\mathbf{F}$ with its optimal solution in Eq. (11), the problem in Eq. (9) w.r.t. $\widetilde{\mathbf{W}}$ is:

$$\begin{aligned}
\min_{\widetilde{\mathbf{W}}} \ &\mathrm{Tr}\big(\mathbf{Q}^T\mathbf{P}^T(\gamma\mathbf{I}+\mathbf{L_S}+\mathbf{U})\mathbf{PQ} - 2\mathbf{Q}^T\mathbf{P}^T(\mathbf{UY}+\gamma\mathbf{X}^T\widetilde{\mathbf{W}})\big) \\
&+ \mathrm{Tr}(\widetilde{\mathbf{W}}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\widetilde{\mathbf{W}}) + \gamma\lambda\sum_{v=1}^{V}\frac{\|\widetilde{\mathbf{W}}_v\|_{2,1}}{\alpha_v},
\end{aligned} \tag{12}$$

where $\mathbf{M} = \gamma\mathbf{I} + 2\beta\mu\mathbf{L_S}$. Due to $\mathbf{Q}^T\mathbf{P}^T(\gamma\mathbf{I}+\mathbf{L_S}+\mathbf{U})\mathbf{PQ} = \mathbf{Q}^T\mathbf{P}^T(\mathbf{UY}+\gamma\mathbf{X}^T\widetilde{\mathbf{W}}) = \mathbf{Q}^T\mathbf{PQ}$, Eq. (12) can be written as:

$$\mathrm{Tr}\Big(\widetilde{\mathbf{W}}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\widetilde{\mathbf{W}} - \mathbf{Q}^T\mathbf{PQ}\Big) + \gamma\lambda\sum_{v=1}^{V}\frac{\|\widetilde{\mathbf{W}}_v\|_{2,1}}{\alpha_v}. \tag{13}$$

By substituting $\mathbf{Q} = \mathbf{UY} + \gamma\mathbf{X}^T\widetilde{\mathbf{W}}$, Eq. (13) is simplified as:

$$\min_{\widetilde{\mathbf{W}}} \mathrm{Tr}\big(\widetilde{\mathbf{W}}^T\mathbf{G}\widetilde{\mathbf{W}}\big) - 2\,\mathrm{Tr}\big(\mathbf{B}^T\widetilde{\mathbf{W}}\big) + \gamma\lambda\sum_{v=1}^{V}\frac{1}{\alpha_v}\sum_{i=1}^{d_v}\|\widetilde{\mathbf{w}}_v^i\|_2, \tag{14}$$

where $\mathbf{G} = \mathbf{X}(\mathbf{M} - \gamma^2\mathbf{P}^T)\mathbf{X}^T$, $\mathbf{B} = \gamma\mathbf{XPUY}$, and $\widetilde{\mathbf{w}}_v^i$ is the i-th row of $\widetilde{\mathbf{W}}_v$. Taking the partial derivative of Eq. (14) w.r.t. $\widetilde{\mathbf{W}}$, and setting it to be zero, we obtain:

$$\mathbf{G}\widetilde{\mathbf{W}} - \mathbf{B} + \gamma\lambda\mathbf{A}\widetilde{\mathbf{W}} = 0 \Longrightarrow \widetilde{\mathbf{W}} = (\mathbf{G} + \gamma\lambda\mathbf{A})^{-1}\mathbf{B}, \tag{15}$$

where $\mathbf{A} = [\mathbf{A}^1, \cdots, \mathbf{A}^V] \in \mathbb{R}^{d\times d}$ is a diagonal matrix with $\mathbf{A}^v = \mathrm{diag}(\frac{\alpha_v^{-1}}{2\|\widetilde{\mathbf{w}}_v^1\|_2}, \cdots, \frac{\alpha_v^{-1}}{2\|\widetilde{\mathbf{w}}_v^{d_v}\|_2}) \in \mathbb{R}^{d_v\times d_v}$. Note that $\mathbf{A}$ is unknown and depends on $\widetilde{\mathbf{W}}$, once $\widetilde{\mathbf{W}}$ is determined, then $\mathbf{A}$ can be updated accordingly. Thus, the optimal solution of $\widetilde{\mathbf{W}}$ can be obtained by updating $\mathbf{A}$ and $\widetilde{\mathbf{W}}$ alternately.

• **Update $\boldsymbol{\alpha}$:** when variables are fixed except $\boldsymbol{\alpha}$, the problem in Eq. (9) w.r.t. $\boldsymbol{\alpha}$ is written as:

$$\min_{\boldsymbol{\alpha}\geq 0, \boldsymbol{\alpha}^T\mathbf{1}=1} \ \sum_{v=1}^{V}\frac{e_v}{\alpha_v}, \tag{16}$$

where $e_v = \|\widetilde{\mathbf{W}}_v\|_{2,1}$. To satisfy the constraints in Eq. (16), the Lagrangian function is formulated as:

$$\mathcal{L}(\boldsymbol{\alpha}, \eta, \boldsymbol{\phi}) = \sum_{v=1}^{V}\frac{e_v}{\alpha_v} + \eta(\boldsymbol{\alpha}^T\mathbf{1} - 1) - \boldsymbol{\phi}^T\boldsymbol{\alpha}, \tag{17}$$

where $\eta \in \mathbb{R}$ and $\boldsymbol{\phi} \in \mathbb{R}^V$ are Lagrangian multipliers. Considering that the optimal $\boldsymbol{\alpha}$ should satisfy the KKT condition [42], we have:

$$\forall v: \begin{cases} -\frac{e_v}{\alpha_v^2} + \eta^* - \phi_v^* = 0, \\ \alpha_v \geq 0, \boldsymbol{\alpha}^T\mathbf{1} = 1, \\ \phi_v^* \geq 0, \phi_v^*\alpha_v = 0, \end{cases} \Rightarrow \alpha_v = \sqrt{\frac{e_v}{\eta^* - \phi_v^*}}, \tag{18}$$

where η^* and ϕ_v^* are the optimal values of η and $\boldsymbol{\phi}$ corresponding to the optimal solution of $\boldsymbol{\alpha}$, respectively. Since $\alpha_v \geq 0$, $\phi_v \geq 0$ and $\phi_v^*\alpha_v = 0$, we know that $\phi_v^* = 0$ if $\alpha_v > 0$, and $\alpha_v = 0$ if $\phi_v^* > 0$. Based on $\alpha^T\mathbf{1} = 1$, we can infer the optimal solution of α

$$\sum_{v=1}^{V}\sqrt{\frac{e_v}{\eta^*}} = 1 \Rightarrow \sqrt{\eta^*} = \sum_{v=1}^{V}\sqrt{e_v} \Rightarrow \alpha_v = \frac{\sqrt{e_v}}{(\sum_{v=1}^{V}\sqrt{e_v})}. \tag{19}$$

The solution of α_v in Eq. (19) satisfies the non-negative constraint because of $e_v \geq 0$ for each v.

• **Update S:** when $\boldsymbol{\alpha}$, $\mathbf{F}$ and $\widetilde{\mathbf{W}}$ are fixed, the optimization subproblem of Eq. (9) becomes:

$$\min_{\mathbf{S1}=\mathbf{1},\mathbf{S}\geq 0} \beta\sum_{v=1}^{V}\|\mathbf{S}-\mathbf{S}^v\|_F + \mathrm{Tr}(\mathbf{F}^T\mathbf{L_S}\mathbf{F} + 2\beta\mu\widetilde{\mathbf{W}}^T\mathbf{X}\mathbf{L_S}\mathbf{X}^T\widetilde{\mathbf{W}}). \tag{20}$$

Algorithm 1 : The optimization algorithm for SMFS

Input: Data $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_V]^T$, ground truth labels $\mathbf{Y}_L$ of labeled data $\mathbf{X}^L$, initialized single-view graphs $\{\mathbf{S}^v\}_{v=1}^V$, and parameters γ, λ, β and μ;

Initialize $\alpha_v = 1/V$ for each view, $\mathbf{S} = \sum_{v=1}^V \mathbf{S}^v/V$, and $\widetilde{\mathbf{W}}$ by least squares regression on $\{\mathbf{X}^L, \mathbf{Y}_L\}$;
repeat
  Updated $\mathbf{F}$ by Eq. (11);
  repeat
    With current $\widetilde{\mathbf{W}}$, update the diagonal matrix $\mathbf{A}$;
    With current $\mathbf{A}$, update $\widetilde{\mathbf{W}}$ by Eq. (15);
  until convergence
  Update $\boldsymbol{\alpha}$ by Eq. (19);
  Update q_v by Eq. (23);
  Update each row of $\mathbf{S}$ by solving Eq. (27);
until the objective function of Eq. (9) converges;

Output: The joint weighted feature selection matrix $\widetilde{\mathbf{W}}$. The matrix is used to calculate the feature scores $\|\widetilde{\mathbf{w}}^i\|_2$ ($i = 1, 2, \cdots, d$), and select p features with the highest scores.

Without introducing explicit weight factors, Eq. (20) implicitly and adaptively fuses multiple single-view graphs $\{\mathbf{S}^v\}_{v=1}^V$ to learn a unified graph $\mathbf{S}$, whose Lagrangian function is:

$$\beta \sum_{v=1}^{V} \|\mathbf{S}-\mathbf{S}^v\|_F + \mathrm{Tr}(\mathbf{F}^T\mathbf{L}_\mathbf{S}\mathbf{F} + 2\beta\mu\widetilde{\mathbf{W}}^T\mathbf{X}\mathbf{L}_\mathbf{S}\mathbf{X}^T\widetilde{\mathbf{W}}) + \Gamma(\mathbf{\Lambda}, \mathbf{S}), \quad (21)$$

where $\mathbf{\Lambda}$ denotes the Lagrange multiplier, and $\Gamma(\mathbf{\Lambda}, \mathbf{S})$ is a formalized term derived from constraints. Taking the derivative of Eq. (21) w.r.t. $\mathbf{S}$ and setting it to be zero, we have

$$\begin{aligned} &\beta \sum_{v=1}^{V} q_v \frac{\partial \|\mathbf{S} - \mathbf{S}^v\|_F^2}{\partial \mathbf{S}} + \frac{\partial \Gamma(\mathbf{\Lambda}, \mathbf{S})}{\partial \mathbf{S}} \\ &+ \frac{\partial \mathrm{Tr}(\mathbf{F}^T\mathbf{L}_\mathbf{S}\mathbf{F} + 2\beta\mu\widetilde{\mathbf{W}}^T\mathbf{X}\mathbf{L}_\mathbf{S}\mathbf{X}^T\widetilde{\mathbf{W}})}{\partial \mathbf{S}} = 0, \end{aligned} \quad (22)$$

where

$$q_v = \frac{1}{2\|\mathbf{S} - \mathbf{S}^v\|_F}. \quad (23)$$

Eq. (22) cannot be directly solved because q_v depends on $\mathbf{S}$. But when q_v is set to be stationary, Eq. (22) essentially boils down to the following problem:

$$\min_{\mathbf{S}\mathbf{1}=\mathbf{1}, \mathbf{S}\geq 0} \beta \sum_{v=1}^{V} q_v \|\mathbf{S}-\mathbf{S}^v\|_F^2 + \mathrm{Tr}(\mathbf{F}^T\mathbf{L}_\mathbf{S}\mathbf{F} + 2\beta\mu\widetilde{\mathbf{W}}^T\mathbf{X}\mathbf{L}_\mathbf{S}\mathbf{X}^T\widetilde{\mathbf{W}}). \quad (24)$$

Thus, Eq. (20) is equivalent to Eq. (24) if q_v is stationary. After we learn $\mathbf{S}$ from Eq. (24), q_v can be determined correspondingly, which inspires us to optimize $\mathbf{S}$ and q_v alternately. Specifically, with fixed q_v, Eq. (24) can be rewritten as:

$$\begin{aligned} \min_{\mathbf{S}\mathbf{1}=\mathbf{1}, \mathbf{S}\geq 0} \frac{1}{2} \sum_{i,j=1}^{n} \|\mathbf{f}^i - \mathbf{f}^j\|_2^2 s_{ij} + \beta \sum_{i=1}^{n} \sum_{v=1}^{V} q_v \|\mathbf{s}_i - \mathbf{s}_i^v\|_2^2 \\ + \beta\mu \sum_{i,j=1}^{n} s_{ij} \|\widetilde{\mathbf{W}}^T\mathbf{x}_i - \widetilde{\mathbf{W}}^T\mathbf{x}_j\|_2^2), \end{aligned} \quad (25)$$

where $\mathbf{f}^i$, $\mathbf{s}_i$ and $\mathbf{s}_i^v$ are the i-th rows of $\mathbf{F}$, $\mathbf{S}$ and $\mathbf{S}^v$, respectively. Note that the optimization problem in Eq. (25) is independent for different i, so that each row of $\mathbf{S}$ (i.e., $\mathbf{s}_i$) can be separately solved as:

$$\min_{\mathbf{s}_i \geq 0, \mathbf{s}_i\mathbf{1}=1} \sum_{v=1}^{V} q_v \|\mathbf{s}_i - \mathbf{s}_i^v\|_2^2 + \mathbf{s}_i\mathbf{d}_i, \quad (26)$$

where $\mathbf{d}_i \in \mathbb{R}^{n\times 1}$ with the j-th element $d_{ij} = \frac{1}{2\beta}\|\mathbf{f}^i - \mathbf{f}^j\|_2^2 + \mu\|\widetilde{\mathbf{W}}^T\mathbf{x}_i - \widetilde{\mathbf{W}}^T\mathbf{x}_j\|_2^2$. By simple algebraic manipulations, Eq. (26) can be further reformulated as:

$$\min_{\mathbf{s}_i \geq 0, \mathbf{s}_i\mathbf{1}=1} \|\mathbf{s}_i - \mathbf{u}_i\|_2^2, \quad (27)$$

where $\mathbf{u}_i = \frac{1}{q}(\sum_{v=1}^V q_v\mathbf{s}_i^v - \mathbf{d}_i/2)$ with $q = \sum_{v=1}^V q_v$. Eq. (27) can now be efficiently optimized with a closed form solution [43], whose Lagrangian function is denoted as:

$$\mathcal{L}(\mathbf{s}_i, \theta_i, \zeta_i) = \frac{1}{2}\|\mathbf{s}_i - \mathbf{u}_i\|_2^2 - \theta_i(s_i\mathbf{1} - 1) - \mathbf{s}_i\zeta_i, \quad (28)$$

where $\theta_i \in \mathbb{R}$ and $\zeta_i \in \mathbb{R}^n$ are Lagrangian multipliers. According to the KKT condition, the optimal solution of $\mathbf{s}_i$ satisfies that:

$$\forall j : \begin{cases} s_{ij} - u_{ij} - \theta_i^* - \zeta_{ij}^* = 0, \\ s_{ij} \geq 0, \mathbf{s}_i\mathbf{1} = 1, \\ \zeta_{ij}^* \geq 0, \zeta_{ij}^* s_{ij} = 0, \end{cases} \quad (29)$$

where θ_i^* and ζ_{ij}^* denote the optimal Lagrangian multipliers of θ_i and ζ_i corresponding to the optimal s_{ij}, respectively. Based on the above constraints, the optimal $s_{ij} = (u_{ij} + \theta_i^*)_+$ if we know θ_i^*, where $x_+ = \max(x, 0)$. Due to $\mathbf{s}_i\mathbf{1} = 1$ and $\zeta_{ij}^* \geq 0$ for $\forall j$, we obtain:

$$\begin{cases} \theta_i^* = t_{ij} - u_{ij} - \bar{\zeta}^*, \\ s_{ij} = t_{ij} - \bar{\zeta}^* + \zeta_{ij}^*, \end{cases} \Rightarrow s_{ij} = (t_{ij} - \bar{\zeta}^*)_+ , \quad (30)$$

where $t_{ij} = u_{ij} - \frac{\mathbf{1}^T\mathbf{u}_i}{n} + \frac{1}{n}$ and $\bar{\zeta}^* = \frac{\mathbf{1}^T\zeta_i^*}{n}$. Thus, we can get s_{ij} if $\bar{\zeta}_i^*$ is known. Based on Eq. (30) and the constraint $s_{ij} \geq 0$ for $\forall j$, we have:

$$\zeta_{ij}^* = (\bar{\zeta}^* - t_{ij})_+ \Rightarrow \bar{\zeta}^* = \frac{1}{n}\sum_{j=1}^{n}(\bar{\zeta}^* - t_{ij})_+ . \quad (31)$$

By introducing a function $f(\bar{\zeta}^*) = \frac{1}{n}\sum_{j=1}^n(\bar{\zeta}^* - t_{ij})_+ - \bar{\zeta}^*$, $\bar{\zeta}^*$ would be achieved when $f(\bar{\zeta}^*) = 0$, which can be iteratively solved with Newton's method, as follows:

$$\bar{\zeta}_{t+1}^* = \bar{\zeta}_t^* - \frac{f(\bar{\zeta}_t^*)}{f'(\bar{\zeta}_t^*)}. \quad (32)$$

SMFS separately solves the subproblems of $\mathbf{F}$, $\widetilde{\mathbf{W}}$, $\boldsymbol{\alpha}$ and $\mathbf{S}$, and repeats these procedures iteratively until the objective function converges. We further summarize the overall pipeline of solving SMFS in Algorithm 1.

III. Algorithm Analysis

This section gives analyses of the proposed SMFS algorithm in three aspects. We first analyze the convergence property of SMFS theoretically, then discuss the relationship between SMFS and some single-view feature selection methods. Finally, we consider the computational complexity of SMFS.

A. Convergence Analysis of SMFS

The variables in SMFS are alternately optimized, since the objective function defined in Eq. (9) is not jointly convex w.r.t. all variables. Therefore, it is necessary to prove that the optimization procedures described in Algorithm 1 can monotonically decrease objective function value in each iteration until convergence. The proof of the convergence of SMFS is: For convenience, suppose that we have $\mathbf{F}_t$, $\widetilde{\mathbf{W}}_t$, $\mathbf{S}_t$, $\boldsymbol{\alpha}^t$ after the t-th iteration, and the corresponding objective function is:

$$\begin{aligned}\mathcal{L}(\mathbf{F}_t, \widetilde{\mathbf{W}}_t, \mathbf{S}_t, \boldsymbol{\alpha}^t) &= \mathrm{Tr}\big(\mathbf{F}_t^T \mathbf{L}_{\mathbf{S}_t} \mathbf{F}_t\big) + \mathrm{Tr}\big((\mathbf{F}_t - \mathbf{Y})^T \mathbf{U} (\mathbf{F}_t - \mathbf{Y})\big) \\ &+ \gamma\Big(\|\mathbf{F}_t - \mathbf{X}^T \widetilde{\mathbf{W}}_t\|_F^2 + \lambda \mathrm{Tr}\big(\widetilde{\mathbf{W}}_t^T \mathbf{A}_t \widetilde{\mathbf{W}}_t\big)\Big) \\ &+ \beta \sum_{v=1}^{V} \|\mathbf{S}_t - \mathbf{S}^v\|_F + 2\beta\mu \mathrm{Tr}\big(\widetilde{\mathbf{W}}_t^T \mathbf{X} \mathbf{L}_{\mathbf{S}_t} \mathbf{X}^T \widetilde{\mathbf{W}}_t\big). \end{aligned} \tag{33}$$

With current $\widetilde{\mathbf{W}}_t$, $\mathbf{S}_t$ and $\boldsymbol{\alpha}^t$, Algorithm 1 obtains $\widetilde{\mathbf{W}}_{t+1}$ in the $t+1$ iteration, which holds that:

$$\mathcal{G}(\widetilde{\mathbf{W}}_{t+1}) + \gamma\lambda \mathrm{Tr}(\widetilde{\mathbf{W}}_{t+1}^T \mathbf{A}_t \widetilde{\mathbf{W}}_{t+1}) \leq \mathcal{G}(\widetilde{\mathbf{W}}_t) + \gamma\lambda \mathrm{Tr}(\widetilde{\mathbf{W}}_t^T \mathbf{A}_t \widetilde{\mathbf{W}}_t), \tag{34}$$

where $\mathcal{G}(\widetilde{\mathbf{W}}) = 2\beta\mu \mathrm{Tr}(\widetilde{\mathbf{W}}^T \mathbf{X} \mathbf{L}_{\mathbf{S}_t} \mathbf{X}^T \widetilde{\mathbf{W}}) + \gamma \|\mathbf{F}_t - \mathbf{X}^T \widetilde{\mathbf{W}}\|_F^2$, $\mathbf{A}_t$ and $\mathbf{L}_{\mathbf{S}_t}$ are respectively computed by $\widetilde{\mathbf{W}}_t$, $\boldsymbol{\alpha}^t$ and $\mathbf{S}_t$. According to the inequality $\sqrt{x} - \frac{x}{2\sqrt{y}} \leq \sqrt{y} - \frac{y}{2\sqrt{y}}$ in [44], it is derived that:

$$\begin{aligned} &\|(\widetilde{\mathbf{W}}_v)_{t+1}\|_{2,1} - \sum_{i=1}^{d_v} \frac{\|(\widetilde{\mathbf{w}}_v^i)_{t+1}\|_2^2}{2\|(\widetilde{\mathbf{w}}_v^i)_t\|_2} \leq \|(\widetilde{\mathbf{W}}_v)_t\|_{2,1} - \sum_{i=1}^{d_v} \frac{\|(\widetilde{\mathbf{w}}_v^i)_t\|_2^2}{2\|(\widetilde{\mathbf{w}}_v^i)_t\|_2} \\ &\Rightarrow \frac{1}{\alpha_v^t}\Big(\|(\widetilde{\mathbf{W}}_v)_{t+1}\|_{2,1} - \sum_{i=1}^{d_v} \frac{\|(\widetilde{\mathbf{w}}_v^i)_{t+1}\|_2^2}{2\|(\widetilde{\mathbf{w}}_v^i)_t\|_2}\Big) \\ &\quad \leq \frac{1}{\alpha_v^t}\Big(\|(\widetilde{\mathbf{W}}_v)_t\|_{2,1} - \sum_{i=1}^{d_v} \frac{\|(\widetilde{\mathbf{w}}_v^i)_t\|_2^2}{2\|(\widetilde{\mathbf{w}}_v^i)_t\|_2}\Big) \\ &\Rightarrow \sum_{v=1}^{V} \frac{\|(\widetilde{\mathbf{W}}_v)_{t+1}\|_{2,1}}{\alpha_v^t} - \mathrm{Tr}(\widetilde{\mathbf{W}}_{t+1}^T \mathbf{A}_t \widetilde{\mathbf{W}}_{t+1}) \\ &\quad \leq \sum_{v=1}^{V} \frac{\|(\widetilde{\mathbf{W}}_v)_t\|_{2,1}}{\alpha_v^t} - \mathrm{Tr}(\widetilde{\mathbf{W}}_t^T \mathbf{A}_t \widetilde{\mathbf{W}}_t), \end{aligned} \tag{35}$$

Summing Eq. (34) with the $\gamma\lambda$ times of Eq. (35), we get:

$$\begin{aligned} &\mathcal{G}(\widetilde{\mathbf{W}}_{t+1}) + \gamma\lambda \sum_{v=1}^{V} \frac{\|(\widetilde{\mathbf{W}}_v)_{t+1}\|_{2,1}}{\alpha_v^t} \leq \mathcal{G}(\widetilde{\mathbf{W}}_t) + \gamma\lambda \sum_{v=1}^{V} \frac{\|(\widetilde{\mathbf{W}}_v)_t\|_{2,1}}{\alpha_v^t} \\ &\Rightarrow \mathcal{L}(\mathbf{F}_t, \widetilde{\mathbf{W}}_{t+1}, \mathbf{S}_t, \boldsymbol{\alpha}^t) \leq \mathcal{L}(\mathbf{F}_t, \widetilde{\mathbf{W}}_t, \mathbf{S}_t, \boldsymbol{\alpha}^t), \end{aligned} \tag{36}$$

which indicates that the objective function decreases monotonically by iteratively optimizing $\widetilde{\mathbf{W}}$ with current $\widetilde{\mathbf{W}}_t$, $\mathbf{S}_t$ and $\boldsymbol{\alpha}^t$. Meanwhile, we note that $\mathbf{F}$ and $\boldsymbol{\alpha}$ can be sequentially updated with the closed-form solutions by fixing $\widetilde{\mathbf{W}}$ as $\widetilde{\mathbf{W}}_{t+1}$. Thus, we directly have:

$$\mathcal{L}(\mathbf{F}_{t+1}, \widetilde{\mathbf{W}}_{t+1}, \mathbf{S}_t, \boldsymbol{\alpha}^{t+1}) \leq \mathcal{L}(\mathbf{F}_t, \widetilde{\mathbf{W}}_{t+1}, \mathbf{S}_t, \boldsymbol{\alpha}^t). \tag{37}$$

Finally, when we fix $\widetilde{\mathbf{W}}$ as $\widetilde{\mathbf{W}}_{t+1}$ and $\mathbf{F}$ as $\mathbf{F}_{t+1}$, SMFS updates $\widetilde{\mathbf{S}}$ with current $\mathbf{S}_t$ and q_v^t, which holds that:

$$\beta \sum_{v=1}^{V} \frac{\|\mathbf{S}_{t+1} - \mathbf{S}^v\|_F^2}{2\|\mathbf{S}_t - \mathbf{S}^v\|_F} + \mathcal{F}(\mathbf{S}_{t+1}) \leq \beta \sum_{v=1}^{V} \frac{\|\mathbf{S}_t - \mathbf{S}^v\|_F^2}{2\|\mathbf{S}_t - \mathbf{S}^v\|_F} + \mathcal{F}(\mathbf{S}_t), \tag{38}$$

where $\mathcal{F}(\mathbf{S}) = 2\beta\mu \mathrm{Tr}(\widetilde{\mathbf{W}}_{t+1}^T \mathbf{X} \mathbf{L}_{\mathbf{S}} \mathbf{X}^T \widetilde{\mathbf{W}}_{t+1}) + \mathrm{Tr}\big(\mathbf{F}_{t+1}^T \mathbf{L}_{\mathbf{S}} \mathbf{F}_{t+1}\big)$. Similarly, for Eq. (35), we can infer that:

$$\begin{aligned} &\sum_{v=1}^{V} \Big(\|\mathbf{S}_{t+1} - \mathbf{S}^v\|_F - \frac{\|\mathbf{S}_{t+1} - \mathbf{S}^v\|_F^2}{2\|\mathbf{S}_t - \mathbf{S}^v\|_F}\Big) \\ &\quad \leq \sum_{v=1}^{V} \Big(\|\mathbf{S}_t - \mathbf{S}^v\|_F - \frac{\|\mathbf{S}_t - \mathbf{S}^v\|_F^2}{2\|\mathbf{S}_t - \mathbf{S}^v\|_F}\Big). \end{aligned} \tag{39}$$

Summing Eq. (38) with the β times of Eq. (39), we get:

$$\begin{aligned} &\beta \sum_{v=1}^{V} \Big(\|\mathbf{S}_{t+1} - \mathbf{S}^v\|_F + \mathcal{F}(\mathbf{S}_{t+1}) \leq \beta \sum_{v=1}^{V} \Big(\|\mathbf{S}_t - \mathbf{S}^v\|_F + \mathcal{F}(\mathbf{S}_t) \\ &\Rightarrow \mathcal{L}(\mathbf{F}_{t+1}, \widetilde{\mathbf{W}}_{t+1}, \mathbf{S}_{t+1}, \boldsymbol{\alpha}^{t+1}) \leq \mathcal{L}(\mathbf{F}_{t+1}, \widetilde{\mathbf{W}}_{t+1}, \mathbf{S}_t, \boldsymbol{\alpha}^{t+1}). \end{aligned} \tag{40}$$

Based on Eq. (36), Eq. (37) and Eq. (40), we can infer that:

$$\mathcal{L}(\mathbf{F}_{t+1}, \widetilde{\mathbf{W}}_{t+1}, \mathbf{S}_{t+1}, \boldsymbol{\alpha}^{t+1}) \leq \mathcal{L}(\mathbf{F}_t, \widetilde{\mathbf{W}}_t, \mathbf{S}_t, \boldsymbol{\alpha}^t). \tag{41}$$

Considering that $\mathcal{L}(\mathbf{F}, \widetilde{\mathbf{W}}, \mathbf{S}, \boldsymbol{\alpha})$ has a lower bound (at least above 0), we conclude that the objective function of SMFS is monotonously decreased until convergence.

B. Connection to Single-View Feature Selection Methods

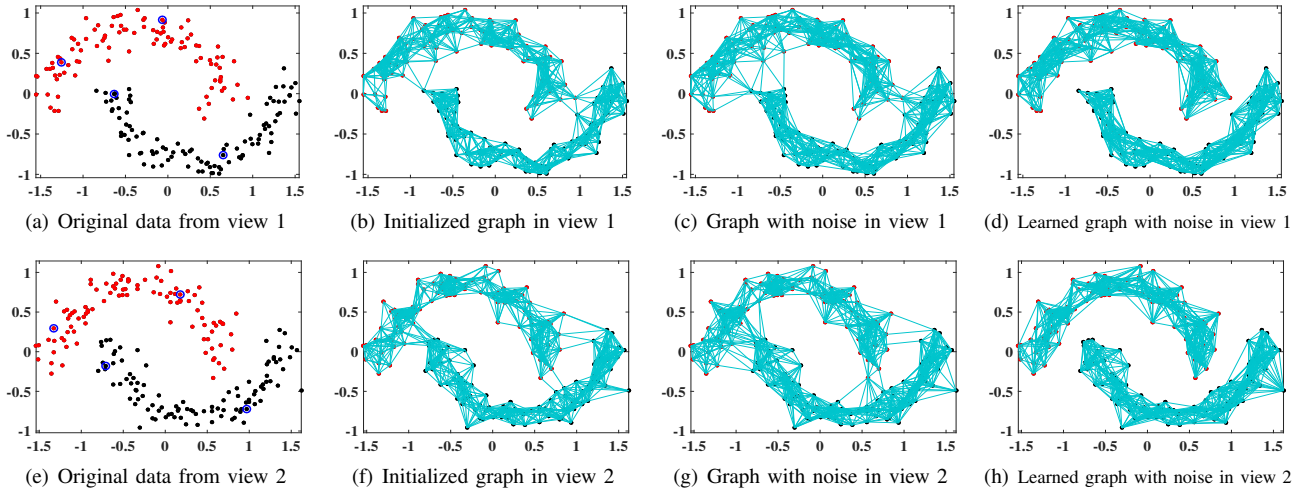
In this subsection, we analyze the relationship between the proposed SMFS and previous single-view feature selection methods. Firstly, when the number of views is set to 1 (i.e., $V = 1$), the proposed SMFS becomes a single-view problem, in which the graph $\mathbf{S}$ is adaptively optimized according to the sample similarity in the original feature space and the projected subspace simultaneously. Accordingly, the single-view version of SMFS, i.e., SSFS, can also be used to validate the effectiveness of SMFS in the next section, and it is formulated as:

$$\begin{aligned} \min_{\mathbf{S1}=\mathbf{1}, \mathbf{S}\geq 0, \mathbf{F}, \mathbf{W}} &\mathrm{Tr}(\mathbf{F}^T \mathbf{L}_{\mathbf{S}} \mathbf{F}) + \mathrm{Tr}\big((\mathbf{F} - \mathbf{Y})^T \mathbf{U} (\mathbf{F} - \mathbf{Y})\big) \\ &+ \beta\big(\|\mathbf{S} - \mathbf{S}^1\|_F^2 + \mu \sum_{i,j=1}^{n} s_{ij} \|\mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j\|_2^2\big) \\ &+ \gamma\big(\|\mathbf{F} - \mathbf{X}^T \mathbf{W}\|_F^2 + \lambda \|\mathbf{W}\|_{2,1}\big). \end{aligned} \tag{42}$$

The optimization of SSFS can use the procedures provided in Algorithm 1. Notably, SSFS is equivalent to the objective function of SFSS [24] if we set $\beta = 0$, and it can be further generalized to an unsupervised method by replacing the label consistency constraint on $\mathbf{F}$ with the orthogonal constraint $\mathbf{F}^T \mathbf{F} = \mathbf{I}$. Similarly, when $\gamma \to \infty$ and $\beta = 0$, SSFS shares the similar goal with [25] by changing $\|\mathbf{W}\|_{2,1}$ to $\|\mathbf{W}\|_{2,p}$ ($p \in (0, 1]$). Thus, these feature selection methods can be viewed as the special cases of the proposed SMFS under the single-view scenario. In another perspective, it can be seen that SMFS performs feature selection and graph learning simultaneously, while most existing methods predefine similarity graph based on the Euclidean distance between samples in the original feature space, and fix the graph during feature selection. Therefore, SMFS can achieve better performance in practice owing to the adaptive multiple graph fusion and learning model for feature selection.

C. Computational Complexity Analysis

The optimization procedure of SMFS is iteratively updating $\mathbf{F}$, $\widetilde{\mathbf{W}}$, $\boldsymbol{\alpha}$ and $\mathbf{S}$ with Eq. (11), Eq. (15), Eq. (19) and Eq. (27). Here, we briefly analyze the computational complexity of SMFS. Specifically, updating $\mathbf{F}$ and $\widetilde{\mathbf{W}}$ involve matrix inversions, and take $\mathcal{O}(n^3)$ and $\mathcal{O}(d^3)$ in each iteration respectively. For the optimization of $\boldsymbol{\alpha}$, it usually takes $\mathcal{O}(dc^2)$ to solve Eq. (19). Since $\mathbf{S}$ is solved with $\mathcal{O}(n)$ for each row, it costs $\mathcal{O}(n^2)$ for entire $\mathbf{S}$. Besides, updating q_v by Eq. (23) is no more than $\mathcal{O}(n^2)$. Due to $c \ll d$ and $c \ll n$,



(a) Original data from view 1 (b) Initialized graph in view 1 (c) Graph with noise in view 1 (d) Learned graph with noise in view 1

(e) Original data from view 2 (f) Initialized graph in view 2 (g) Graph with noise in view 2 (h) Learned graph with noise in view 2

Fig. 1. The similarity graph learned by SMFS and the initialized graphs with or without noise for different views of the Two-moon dataset.

the total computational complexity of SMFS is approximated by $\mathcal{O}(Td^3 + Tn^3)$, where T is the number of iterations in Algorithm 1. It is notable that T is less than 5 for SMFS to converge empirically (see Section IV-E).

IV. Experiments

In this section, we present the experimental studies of the proposed SMFS. Firstly, we visually demonstrate the ability of SMFS in capturing the similarity structure of data, and further quantify the impact of noisy features (or views) on the performance of SMFS. Secondly, we perform experiments on real multi-view datasets to further verify the superiority of SMFS over state-of-the-art methods. Furthermore, experiments on emotion recognition datasets are conducted to intuitively demonstrate the effectiveness of SMFS in terms of identifying critical views among multiple views and selecting discriminative features within each view. Finally, we show the parameter sensitivity and convergence analysis of SMFS.

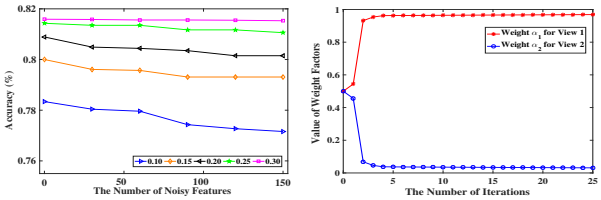
A. Experiments on Synthetic Dataset

We followed [32] to randomly generate a synthetic dataset (e.g., Two-moon) to demonstrate the capability of SMFS for multi-view data. The Two-moon dataset consists of two views from 2 clusters, in which each cluster contains 100 sample points. The first two dimensions of both views are generated with a moon pattern with 0.12 percentage noise and are shown in Figs. 1(a) and 1(e) respectively, where each moon stands for one cluster and the points marked with blue circles denote the labeled samples. To intuitively verify the effectiveness of SMFS, 8 additional noisy features randomly distributed in the range of -0.2 to 0.2 are also added to each view of Two-moon data, resulting in totally 10 features in each view. Figs. 1(b) and 1(f) depict the initialized similarity graphs that are directly constructed from the two-dimension feature space, and Figs. 1(c) and 1(g) show the graphs that are built from the feature space containing the noisy features. From the results, it is evident that although the graphs in Figs. 1(b) and 1(f) can roughly identify the data distribution structure, different clusters are still connected with several lines. Furthermore, the connecting lines between the two clusters in Figs. 1(c) and 1(g) are significantly strengthened when the noisy features are added to the feature space. This visualization indicates that the noisy features adversely affect the sample similarity structure and degrade the reliability of graph, thereby making the clusters even harder to be separated. Figs. 1(d) and 1(h) show the similarity graphs that are dynamically learned from the original feature space containing noise as well as the projected feature subspace. From these figures, we observe that the learned graphs correctly explore the data distribution structure, such that different clusters can be explicitly separated without any connecting lines, which means the similarity graph learned by Eq. (7) can deliver more discriminant information and thus is more effective for label propagation and feature selection than the initialized single-view graphs. This indicates that SMFS can exploit the similarity structure in projected feature space to learn a reliable graph across multiple views, even though there exist noisy features.

B. Experiments on Waveform Dataset

To further analyze the impact of noisy features and views on the performance of SMFS, we conducted experiments on the Waveform[1] dataset, which is widely used in feature selection tasks [45], [46]. The Waveform contains $5,000$ samples with 40 features, consisting of 21 normal features (used as the first view) and 19 noisy features (used as the second view). Obviously, the noisy features in the second view would hamper feature selection, such that an ideal multi-view feature selection should identify the relevant features from the first view and simultaneously assign appropriate weight to each view. In this experiment, the Waveform is randomly partitioned into 2 parts, i.e., 1000 samples for training and the remaining for testing. To verify SMFS, we add noisy features to the above two views gradually and evenly, in which the noisy features are independent and identically distributed from $\mathcal{N}(0, 1)$. SMFS is firstly run 20 times on disparate training sets to select

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/waveform/



(a) Accuracies versus noisy features (b) Allocated weights for two views

Fig. 2. The experimental results on Waveform, in which (a) shows the accuracy variation curves with gradually added noisy features, and (b) shows the view weights with respect to the number of iterations.

optimal feature subsets, in which the proportion of labeled samples varies from 10% to 30%, then the top 10 features selected by SMFS are adopted to train the regularized least-squares classifier (RLSC). The average classification accuracy on different testing sets is used to evaluate the effectiveness of SMFS on feature selection.

To quantify the impact of noisy features on performance more intuitively, the accuracies against the gradually increasing noisy features are illustrated in Fig. 2(a). As can be seen, the accuracy with different proportions of labeled samples has a slightly decreased trend with the increase of noisy features, demonstrating better robustness against noisy features. It is particularly noteworthy that SMFS behaves more stably against the increase of noisy features when the labeled proportion is larger than 15%. This indicates that SMFS is more immune to the noisy features with more labeled samples. Furthermore, Fig. 2(b) records the view weights after each iteration when the labeled proportion and the number of noisy features are 20% and 60, respectively. From the results in Fig. 2, we observe that: **a)** SMFS can identify the relevant features for classification from the first view such that the performance is less affected by the noisy features; **b)** the noisy features in the second view likewise cannot degrade the performance since SMFS assigns appropriate weights to different views, thereby weakening its role. Therefore, SMFS maintains promising effectiveness for multi-view feature selection in the presence of noisy features and views.

C. Experiments on Multi-view Datasets

1) Experimental Setting: Towards the further evaluation of the proposed SMFS algorithm, we employ six real multi-view benchmark datasets including MSRC-v1, Handwritten (HW), Caltech101-7 (Cal-7), COIL20, ORL and SCENE. The details about all datasets including the numbers of classes, the data sizes, and the dimension of features (i.e., the values in brackets) in different views are summarized in Table I.

To comprehensively verify the effectiveness and superiority of SMFS, we compare SMFS with different feature selection methods, including four semi-supervised single-view methods and four state-of-the-art multi-view feature selection methods, respectively. We briefly introduce these methods below:

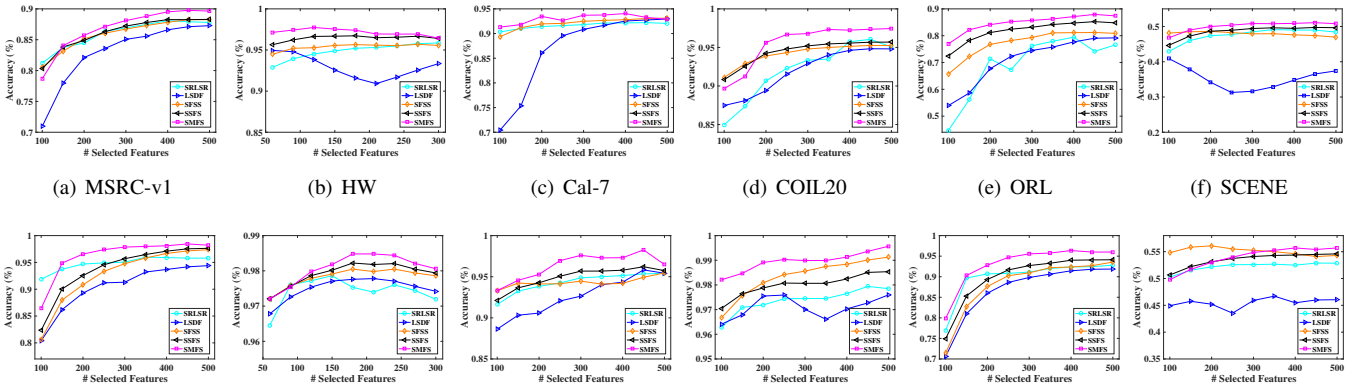
- Locality sensitive discriminant feature (LSDF) [21] is a single-view method that selects features based on the label information and the distribution structure of data.

TABLE I
THE DETAILED INFORMATION OF MULTI-VIEW DATASETS.

View	MSRC-v1	HW	Cal-7
#1	CENTRIST(1302)	PIX(240)	GABOR(48))
#2	CMT(48)	FOU(76)	WM(40)
#3	GIST(512)	FAC(216)	CENTRIST(254)
#4	HOG(100)	ZER(47)	HOG(1984)
#5	LBP(256)	KAR(64)	GIST(512)
#6	SIFT(200)	MOR(6)	LBP(928)
Feature size	2418	649	3766
Classes	7	40	7
Data size	210	2000	1474
View	COIL20	ORL	SCENE
#1	GIST(512)	GIST(512)	GIST(512)
#2	HOG(420)	LBP(59)	CM(432)
#3	LBP(1239)	HOG(864)	HOG(256)
#4	SIFT(630)	CENTRIST(254)	LBP(48)
Feature size	2801	1689	1248
Classes	20	40	8
Data size	1440	400	2688

- Structural feature selection with sparsity (SFSS) [24] is a state-of-the-art single-view semi-supervised method that incorporates sparse l_{21}-norm into manifold regularization.
- Sparse Rescaled Linear Square Regression (SRLSR) [25] is a single-view method without a similarity graph.
- SSFS is the single-view version of SMFS, which directly employs the concatenated features to learn a similarity graph for feature selection. SSFS is used to verify the effectiveness of the proposed feature selection and graph learning model for multi-view data.
- Multi-view sparse feature selection (MSFS) [47] is a supervised method, which validates if the introduction of unlabeled data improves the effectiveness of semi-supervised multi-view feature selection.
- MRMVFS [10] is a semi-supervised multi-view method that integrates the label information, data distribution and correlation among multiple views to select features by applying manifold regularization into multi-view scenario.
- MLSFS [26] exploits multi-view Laplacian regularization to extend the single-view semi-supervised method and replaces the l_{21}-norm by $l_{2,1/2}$-norm for feature selection.
- MASFS [30] is a semi-supervised multi-view feature selection method that adaptively changes the single-view similarity graph based on the sample distance information derived from the original feature space and the feedback information of the current prediction label.

We randomly partition each dataset into 2 subsets, i.e., 80% samples for training and the remaining samples for testing, in which each training set is also randomly divided into the labeled samples and unlabeled samples according to different labeled ratios. For a fair comparison, we tune the parameters of all competitors in the same way as given in the respective research works [21], [24], [48]. The regularization parameters of SFMS are searched in a grid of $\{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$. Considering that different datasets have different dimensions of features, the numbers of selected features vary from $\{60, 90, \cdots, 300\}$ for HW and $\{100, 150, \cdots, 500\}$ for other datasets. Following the semi-supervised feature selection studies [49], in the first step, each method is implemented on the training set to select the optimal feature subset, then the



(a) MSRC-v1 (b) HW (c) Cal-7 (d) COIL20 (e) ORL (f) SCENE

(g) MSRC-v1 (h) HW (i) Cal-7 (j) COIL20 (k) ORL (l) SCENE

Fig. 3. The classification accuracy of SMFS and the single-view feature methods on concatenated features with the different numbers of selected features. The figures in the first and second rows display the experimental results with 10% and 30% labeled samples, respectively.

RLSC with a fixed regularization parameter (i.e., $C = 0.1$) is adopted to train a classifier based on the selected features of labeled samples. The trained classifier is used to assess the classification accuracy on testing samples with selected features. We use classification accuracy as a measure to evaluate the effectiveness of selected features in this paper. To reduce the statistical variability, each feature selection method is independently run 20 times on disparate training and testing partitions and the average classification accuracy with the optimal parameter configuration is reported.

2) Experimental Results: ***Comparison to Single-view Methods.*** To evaluate the quality of selected features, the proposed SMFS is firstly compared with single-view methods which directly concatenate multi-view features as single-view data for feature selection. The experiment is performed with 10% and 30% labeled samples, respectively. We measure the classification accuracy of SMFS and the single-view algorithms (i.e., SRLSR, LSDF, SFSS, SSFS) on the benchmark datasets with different numbers of selected features and show the results in Fig. 3. We make the following observation

- As the number of selected features increases, the accuracy of SMFS tends to increase gradually and is considerably superior to the single-view methods that directly use the concatenated features in most cases. Guided by the feature selection model in Eq. (6), SMFS can effectively exploit the consensus and complementarity information contained in multiple views by assigning adaptive weights for different views, facilitating the selection of discriminative features. This also indicates that it is inappropriate to treat each view equally and indiscriminately concatenate multiple projections for joint feature selection.
- SSFS (i.e., the single-view version of SMFS) always outperforms the single-view method SFSS which regards the construction of graph and feature selection as two individual parts and keeps the sample similarity information extracted from the original feature space unchanged during the feature selection process. It verifies that the implicit graph fusion and dynamical similarity learning indeed improve the quality of selected features.

Consequently, compared to the single-view methods that directly select features from the concatenated features of multiple views, SMFS can discriminate different views with adaptive weights and coalesce them to learn a joint weighted feature projection across multiple views, which is more effective in identifying informative and discriminative features.

Comparison to Multi-view Methods. To further validate the effectiveness of the newly proposed SMFS, we compare the performance of SMFS and four state-of-the-art multi-view feature selection methods with 10% and 30% labeled samples. Table II reports the classification accuracy computed by the RLSC with a varying number of selected features. The last column of Table II records the average and the standard deviation of using the different numbers of features for each feature selection method. From the experimental results in Table II, we make the following observations

- As the number of selected features increases, the accuracies of all multi-view methods raise. Furthermore, SMFS achieves competitive or significantly better results on all datasets than the other competitors, fully showing its effectiveness for multi-view feature selection.
- The accuracy using features selected by SMFS is considerably superior to that of the supervised MSFS, which validates that learning the similarity structure of unlabeled samples benefits the selection of more informative features and accordingly achieves better performance.
- Compared with the semi-supervised methods MRMVFS, MLSFS and MASFS, the proposed SMFS shows better or highly competitive performance with the different numbers of selected features. This indicates that it is effective to adaptively optimize graph structure according to the sample similarity both in the original and projected space and to learn a joint feature projection with adaptive weights across all views for multi-view feature selection.

Therefore, SMFS achieves significant competitiveness in comparison with the state-of-the-art multi-view feature selection methods, validating the effectiveness and superiority of the multi-view feature selection in Eq. (6) equipped with the adaptive multiple graph fusion and learning.

To verify the efficiency of SMFS, we implement each feature selection method in the same computing environment

TABLE II
THE CLASSIFICATION ACCURACY (%) WITH VARYING NUMBERS OF SELECTED FEATURES BY SMFS AND OTHER MULTI-VIEW FEATURE SELECTION METHODS WHEN 10% AND 30% DATA ARE RESPECTIVELY LABELED ON THE BENCHMARK DATASETS. THE BEST RESULTS ARE IN BOLD.

Dataset	MSRC-v1 with 10% labeled data										MSRC-v1 with 30% labeled data									
Method	100	150	200	250	300	350	400	450	500	Average	100	150	200	250	300	350	400	450	500	Average
MSFS	69.3	72.0	74.3	74.8	75.5	76.7	77.5	78.0	78.6	75.2±3.0	76.9	78.8	84.8	88.8	91.9	93.2	93.6	93.5	95.0	88.5±6.8
MLSFS	73.2	80.4	82.7	83.8	84.4	85.8	85.8	86.1	86.7	83.2±4.2	85.6	89.9	93.0	93.2	93.7	96.0	96.3	96.4	96.8	93.4±3.7
MRMVFS	73.0	79.6	81.9	84.2	83.8	84.9	86.1	85.4	87.1	83.0±4.3	85.7	92.1	93.9	94.9	96.2	96.7	96.8	97.3	97.0	94.5±3.7
MASFS	74.7	80.5	83.1	85.2	85.6	86.4	87.0	87.4	87.7	84.2±4.2	82.4	90.4	94.3	95.4	97.4	97.5	97.0	97.4	97.5	94.3±5.1
SMFS	**78.7**	**84.1**	**85.7**	**87.1**	**88.1**	**88.8**	**89.5**	**89.8**	**89.6**	**86.8±3.6**	**86.4**	**94.9**	**96.6**	**97.4**	**97.9**	**98.0**	**98.1**	**98.5**	**98.2**	**96.0±3.8**
Dataset	HW with 10% labeled data										HW with 30% labeled data									
Method	60	90	120	150	180	210	240	270	300	Average	60	90	120	150	180	210	240	270	300	Average
MSFS	88.8	91.2	92.4	93.0	93.4	93.8	93.9	94.0	94.1	92.7±2.8	92.7	93.9	94.7	95.1	95.5	95.8	95.9	95.9	96.1	95.1±1.1
MLSFS	94.0	94.5	95.1	95.5	95.7	96.0	96.1	96.3	**96.4**	95.5±0.8	97.5	**97.8**	**98.0**	98.1	98.0	98.0	97.9	97.8	97.8	97.9±0.2
MRMVFS	93.5	94.5	95.0	95.4	95.8	95.9	96.1	96.2	**96.4**	95.4±0.9	96.5	97.2	97.4	97.5	97.8	97.9	98.1	98.0	98.0	97.6±0.5
MASFS	95.7	96.1	96.1	96.1	96.2	96.2	96.2	96.2	96.2	96.1±0.2	**97.6**	**97.8**	97.9	98.1	98.0	98.0	98.0	98.0	97.8	97.9±0.1
SMFS	**96.6**	**96.9**	**97.2**	**97.0**	**96.9**	**96.5**	**96.4**	**96.4**	95.9	**96.6±0.4**	97.2	97.5	**98.0**	**98.2**	**98.5**	**98.5**	**98.4**	**98.2**	**98.1**	**98.1±0.4**
Dataset	Cal-7 with 10% labeled data										Cal-7 with 30% labeled data									
Method	100	150	200	250	300	350	400	450	500	Average	100	150	200	250	300	350	400	450	500	Average
MSFS	84.1	86.9	87.6	88.6	90.1	90.9	91.3	91.8	92.1	89.3±2.7	91.9	93.6	94.5	94.9	95.2	95.5	95.6	96.0	95.8	94.8±1.3
MLSFS	88.7	90.0	90.1	90.9	91.3	91.4	91.6	91.8	92.1	90.9±1.1	91.8	93.4	94.5	95.2	95.4	95.7	96.0	96.1	96.3	94.9±1.5
MRMVFS	**91.3**	**92.0**	92.3	92.6	92.8	92.9	92.8	92.9	92.9	92.5±0.5	92.8	94.2	94.9	95.5	95.6	96.0	96.0	96.2	96.4	95.3±1.1
MASFS	90.5	92.1	92.5	92.6	92.9	93.0	93.0	93.0	**93.0**	92.5±0.8	92.4	94.2	**95.8**	96.3	96.6	96.7	96.7	96.7	**96.7**	95.8±1.5
SMFS	**91.3**	91.8	**93.5**	**92.7**	**93.7**	**93.8**	**94.1**	**93.3**	92.9	**93.0±0.9**	**93.3**	**94.6**	95.3	**97.0**	**97.6**	**97.3**	**97.3**	**98.3**	96.5	**96.3±1.6**
Dataset	COIL20 with 10% labeled data										COIL20 with 30% labeled data									
Method	100	150	200	250	300	350	400	450	500	Average	100	150	200	250	300	350	400	450	500	Average
MSFS	68.8	71.8	85.7	88.8	90.8	91.8	92.7	93.3	93.6	86.4±9.5	90.4	93.1	94.3	95.2	96.1	96.5	96.8	97.0	97.3	95.2±2.3
MLSFS	86.8	89.7	92.2	93.3	94.2	94.7	94.8	95.1	95.0	92.9±2.9	97.4	97.8	98.3	97.8	97.7	98.1	98.3	98.7	98.9	98.1±0.5
MRMVFS	**90.2**	**92.0**	94.0	94.4	94.6	94.9	95.1	95.3	95.5	94.0±1.8	95.0	96.2	97.0	97.7	97.9	98.3	98.5	98.7	98.8	97.6±1.3
MASFS	88.9	90.9	93.4	94.3	94.8	95.3	95.7	95.7	95.8	93.8±2.4	96.8	97.9	98.3	98.6	98.8	98.9	98.9	99.1	99.1	98.5±0.7
SMFS	89.7	91.2	**95.6**	**96.7**	**96.8**	**97.3**	**97.2**	**97.4**	**97.5**	**95.5±2.9**	**98.2**	**98.5**	**98.9**	**99.0**	**99.0**	**99.0**	**99.1**	**99.4**	**99.6**	**98.9±0.4**
Dataset	ORL with 10% labeled data										ORL with 30% labeled data									
Method	100	150	200	250	300	350	400	450	500	Average	100	150	200	250	300	350	400	450	500	Average
MSFS	55.5	59.3	62.8	64.6	65.9	67.6	68.6	69.1	70.2	64.8±4.9	77.4	80.8	82.8	83.8	84.4	85.6	86.7	87.4	88.3	84.1±3.4
MLSFS	61.5	67.3	70.2	72.0	74.1	75.5	77.9	78.1	78.7	72.8±5.7	84.6	89.1	89.7	90.9	91.1	91.6	91.9	92.5	92.9	90.5±2.5
MRMVFS	72.1	76.9	78.6	79.9	81.4	81.8	82.9	83.6	83.9	80.1±3.8	77.6	82.1	87.5	90.6	91.9	92.8	93.6	93.8	94.1	89.3±5.8
MASFS	71.4	78.1	80.3	81.8	82.0	82.4	82.9	83.4	83.5	80.6±3.9	**86.6**	88.8	89.9	90.6	91.6	91.5	92.1	92.1	92.3	90.6±1.9
SMFS	**76.9**	**82.2**	**84.1**	**85.3**	**85.7**	**86.3**	**87.1**	**87.9**	**87.4**	**84.8±3.4**	79.9	**90.3**	**92.8**	**94.7**	**95.6**	**95.8**	**96.4**	**96.0**	**96.0**	**92.7±5.3**
Dataset	SCENE with 10% labeled data										SCENE with 30% labeled data									
Method	100	150	200	250	300	350	400	450	500	Average	100	150	200	250	300	350	400	450	500	Average
MSFS	46 .8	47.8	47.6	47.5	47.3	47.0	46.9	46.5	46.3	47.1±0.5	52.0	52.9	53.0	52.2	52.5	52.1	51.8	51.6	51.4	52.2±0.6
MLSFS	42.5	45.2	47.0	47.7	48.5	48.6	48.8	48.7	48.6	47.3±2.1	51.1	52.5	52.3	53.2	53.4	53.4	53.1	53.5	53.3	52.9±0.8
MRMVFS	46.2	47.9	48.7	48.8	49.3	49.0	49.0	48.6	48.5	48.4±0.9	53.8	54.1	**54.5**	**55.0**	53.9	53.4	53.1	53.1	52.7	53.8±0.7
MASFS	**50.1**	**50.4**	**50.1**	49.8	49.3	49.6	49.3	48.9	48.7	49.6±0.6	**54.0**	**54.3**	54.2	54.0	53.9	53.8	53.5	53.7	53.3	**53.9±0.3**
SMFS	46.8	49.0	50.0	**50.4**	**50.8**	**50.8**	**50.9**	**51.1**	**50.8**	**50.1±1.4**	49.8	51.7	53.0	54.0	**54.9**	**55.3**	**55.7**	**55.4**	**55.7**	53.7±2.1

TABLE III
THE RUNNING TIME (MEASURED BY SECOND) OF SMFS AND OTHER MULTI-VIEW FEATURE SELECTION METHODS ON MULTI-VIEW DATASETS.

Dataset	MSRC-v1	HW	Cal-7	COIL-20	ORL	SCENE
MSFS	1.07	0.35	3.37	2.65	0.69	0.51
MLSFS	19.55	5.26	30.98	15.98	4.06	9.98
MRMVFS	12.91	3.47	18.10	12.51	2.99	8.24
MASFS	13.25	5.83	26.34	18.74	4.58	12.51
SMFS	7.07	3.98	11.67	10.69	2.11	7.27

and record their CPU running time on six multi-view datasets. The average running time of five multi-view methods using 10% labeled samples are reported in Table III. From Table III, it can be observed that the running time varies according to the characteristics of dataset (e.g., the data size and the feature size), and the supervised MSFS always performs more efficiently than semi-supervised methods since MSFS only uses few labeled samples. For all semi-supervised methods, our SMFS achieves better results on five multi-view datasets. On dataset HW, SMFS consumes less time than other methods, except for MRMVFS. The results show that the computational efficiency of SMFS is comparable or considerably superior to the state-of-the-art competitors. The reason is that although SMFS involves matrix inversions, its objective function converges rapidly, which has been theoretically proved in Section III-A and will be demonstrated experimentally in Section IV-E.

Ablation study. Here, we conduct an ablation study to further illustrate the significance of the adaptive weights of feature projections and the proposed graph fusion and learning model. We first remove the procedures for optimizing the view weights $\{\alpha_v\}_{v=1}^V$ and set each α_v to be the average number of views, i.e., $1/V$, thus getting a simplified version of SMFS (named SMFS$_0$) that treats different projections equally. Then, we eliminate the adaptive graph fusion and learning part of SMFS to get another version of SMFS (named SMFS$_1$), which uses a fixed graph (i.e., $\mathbf{S} = \sum_{v=1}^V \mathbf{S}^v/V$) during the process of feature selection. From the results in Table IV, we observe that the performance of SMFS$_0$ is consistently inferior to those of SMFS with different numbers of selected features. This indicates that the low-quality projections will adversely affect feature selection if different feature projections are assigned with the same weights, making the selected features unreliable consequently. With the adaptive weights $\{\alpha_v\}_{v=1}^V$, SMFS can effectively discriminate and fuse different feature projections, not only considering the complementarity among projections but also facilitating the feature selection. Meanwhile, SMFS is considerably superior to SMFS$_1$ that uses a fixed graph. This verifies that the proposed graph fusion

TABLE IV
THE CLASSIFICATION ACCURACY (%) OF SMFS, $SMFS_0$ AND $SMFS_1$ ON ALL BENCHMARK DATASETS.

		10% labeled data with different numbers of selected features										30% labeled data with different numbers of selected features									
Dataset	Method	100	150	200	250	300	350	400	450	500	Average	100	150	200	250	300	350	400	450	500	Average
MSRC-v1	SMFS	78.7	84.1	85.7	87.1	88.1	88.8	89.5	89.8	89.6	86.8±3.6	86.4	94.9	96.6	97.4	97.9	98.0	98.1	98.5	98.2	96.0±3.8
	$SMFS_0$	72.3	79.2	83.5	86.0	87.6	86.7	85.5	85.7	85.7	83.6±4.8	81.7	91.4	93.6	92.9	95.2	95.0	95.0	96.9	97.0	93.2±4.6
	$SMFS_1$	66.7	72.0	76.1	78.9	80.2	81.8	82.9	83.9	83.6	78.5±5.9	76.0	86.8	90.6	93.3	93.7	94.9	96.4	96.7	96.7	92.8±4.1
Cal-7	SMFS	91.3	91.8	93.5	92.7	93.7	93.8	94.1	93.3	92.9	93.0±0.9	93.3	94.6	95.3	97.0	97.6	97.3	97.3	98.3	96.5	96.3±1.6
	$SMFS_0$	87.3	90.2	92.8	91.5	91.9	92.4	92.0	91.7	91.3	91.2±1.6	91.1	94.3	95.3	95.3	95.5	96.1	96.6	96.5	95.5	95.1±1.7
	$SMFS_1$	90.8	92.1	92.0	91.3	89.7	86.4	89.8	91.5	92.3	90.6±1.9	91.8	93.1	94.0	94.0	95.1	95.7	96.1	96.3	95.3	94.7±1.5
COIL20	SMFS	89.7	91.2	95.6	96.7	96.8	97.3	97.2	97.4	97.5	95.5±2.9	98.2	98.5	98.9	99.0	99.0	99.0	99.1	99.4	99.6	98.9±0.4
	$SMFS_0$	82.2	87.4	88.4	92.6	95.0	95.9	96.2	96.3	96.3	92.6±4.4	96.8	97.5	97.8	97.4	96.9	96.3	97.9	98.5	99.0	97.6±0.9
	$SMFS_1$	90.5	92.1	94.0	95.0	95.3	95.4	95.7	95.8	96.0	94.4±1.9	95.4	96.0	96.3	96.3	96.4	96.4	96.7	96.9	97.1	96.4±0.5
ORL	SMFS	76.9	82.2	84.1	85.3	85.7	86.3	87.1	87.9	87.4	84.8±3.4	79.9	90.3	92.8	94.7	95.6	95.8	96.4	96.0	96.0	92.7±5.3
	$SMFS_0$	68.4	75.5	78.3	81.5	81.5	82.3	83.4	84.0	83.7	79.8±5.1	74.8	86.5	91.0	93.1	93.4	94.5	93.9	94.5	95.0	90.7±6.5
	$SMFS_1$	71.8	78.3	80.9	81.9	82.9	83.1	82.8	82.9	82.7	80.8±3.7	73.0	85.3	92.3	93.1	93.6	93.8	94.5	94.6	94.9	90.6±7.2
SCENE	SMFS	46.8	49.0	50.0	50.4	50.8	50.8	50.9	51.1	50.8	50.1±1.4	49.8	51.7	53.0	54.0	54.9	55.3	55.7	55.4	55.7	53.7±2.1
	$SMFS_0$	43.3	45.7	47.9	48.2	50.2	48.8	46.5	48.7	48.9	47.6±2.1	47.4	50.1	50.6	51.2	52.2	53.2	53.7	52.4	53.0	51.5±2.0
	$SMFS_1$	43.4	45.6	46.7	47.2	47.8	47.7	47.8	48.0	47.9	46.9±1.5	52.7	51.8	50.8	50.1	49.4	48.5	47.6	46.6	46.0	49.3±2.3
Dataset	Method	60	90	120	150	180	210	240	270	300	Average	60	90	120	150	180	210	240	270	300	Average
HW	SMFS	96.6	96.9	97.2	97.0	96.9	96.5	96.4	96.4	95.9	96.6±0.4	97.2	97.5	98.0	98.2	98.5	98.5	98.4	98.2	98.1	98.1±0.4
	$SMFS_0$	95.4	95.7	95.2	93.9	92.9	91.9	91.8	92.5	92.8	93.6±1.5	96.4	97.0	98.0	98.2	98.0	97.7	97.4	97.4	97.3	97.5±0.6
	$SMFS_1$	95.4	95.2	94.9	94.6	94.1	93.9	93.9	94.3	94.4	94.5±0.6	96.0	96.4	96.0	96.8	96.7	96.7	96.5	96.4	96.4	96.4±0.4

TABLE V
THE ACCURACY (MEAN%±STD %) OF USING TOP 20 FEATURES SELECTED BY DIFFERENT MULTI-VIEW FEATURE SELECTION METHODS WHEN WE RANDOMLY SELECT 10 LABELED SAMPLES IN EACH CLASS ON THE SEED DATASET. THE BEST RESULTS ARE IN BOLD, AND * DENOTES THAT THE RESULT IS NOT SIGNIFICANTLY WORSE THAN THE BEST USING THE PAIRED T-TEST AT THE 5% SIGNIFICANCE LEVEL.

Subject	1	2	3	4	5	6	7	8	9	10	11	Average
LSDF	88.7±3.8	84.8±3.3	91.2±3.5	93.7±0.5	91.7±1.4	92.6±4.7	88.5±1.0	88.0±3.0	92.4±3.8	93.8±1.4	83.9±4.2	89.3±3.4
SFSS	90.7±2.9	89.4±1.3	94.1±1.5	95.8±0.7	93.5±2.3	93.3±4.4	93.0±2.9	90.1±0.9	93.3±3.7	94.7±0.5	90.6±1.1	92.6±2.1
SRLSR	91.6±0.9	91.4±2.0	93.7±1.8	93.5±1.5	92.2±1.6	93.8±3.9	91.6±1.0	91.1±1.6	92.1±4.2	93.7±0.7	91.7±1.7	92.5±1.2
SSFS	91.3±1.3	91.0±2.2	94.7±1.8	96.3±1.3	93.7±2.0	94.2±2.7	93.9±0.8	91.9±1.4	94.0±3.0	94.4±2.2	91.4±3.0	93.3±1.7
MSFS	89.8±3.5	88.5±2.3	92.9±2.5	95.8±0.6	93.0±1.1	91.3±3.7	91.2±1.1	88.8±0.8	92.6±3.5	92.8±0.7	89.6±2.4	91.5±2.2
MLSFS	90.5±3.0	89.9±2.3	94.3±2.7	96.3±0.2	94.2±1.5	93.7±4.7*	92.5±0.7	91.1±1.3	93.6±1.6	96.4±1.2	90.2±2.8	93.0±2.3
MRMVFS	93.0±1.0*	92.7±2.8	94.3±3.0*	96.0±2.3	93.9±2.2	94.6±4.2*	94.4±1.4	92.8±2.5	94.4±2.3	94.7±0.8	93.5±1.6	94.0±1.0
MASFS	92.0±2.6	92.1±1.6	95.2±2.3*	96.9±0.6	94.8±0.6	94.3±4.3*	93.8±0.7	92.3±1.3	95.1±2.2	96.8±0.5	92.0±0.7	94.1±1.8
SMFS	**95.4±3.1**	**96.6±1.8**	**97.8±1.8**	**99.3±0.4**	**97.4±1.3**	**97.2±2.7**	**97.3±0.9**	**97.1±1.3**	**98.5±1.0**	**98.9±0.1**	**96.8±1.0**	**97.5±1.1**

and learning model indeed enhances the performance of multi-view feature selection. With the implicit graph fusion and the adaptive structure learning in projection subspace, SFMS can accurately capture the similarity structures across multiple views, improving the reliability of label prediction and thus identifying more discriminative features.

D. Applications on Emotion Recognition Datasets

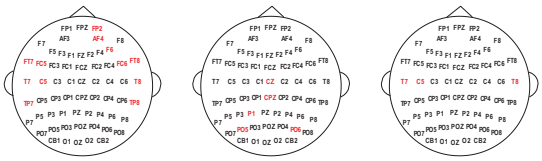
In this section, we apply SMFS to emotion recognition on a newly developed emotional EEG dataset, i.e., the SJTU Emotion EEG dataset (SEED)[2]. The SEED contains the EEG neural signals of multiple subjects who participated in stimulation experiments, where each subject watched 15 emotional film clips that lasted 57 minutes in total, and their emotional reactions are used as the emotional labels of the corresponding film clips. Each subject joined the stimulation experiments in three separate sessions, and each session generated 3394 samples (1120 negative samples, 1104 neutral samples, and 1170 positive samples). The neural signals were recorded at a sampling rate of 200 Hz and processed by a frequency filter with five frequency bands, i.e., Delta (1–3 Hz), Theta (4–7 Hz), Alpha (8–13 Hz), Beta (14–30 Hz), and Gamma (31–50 Hz). Following [50], the differential entropy features extracted from the above frequency bands are chosen for emotion recognition. Hence, each frequency band has 62 electrode features over all the brain areas (shown in Fig. 4) and can be used as one view, so SEED can be regarded as a multi-view dataset with 5 views and 310 features in total. We randomly take 10 samples with label information per class from each session of one subject and 30% unlabeled samples as the training set, and the remaining samples are used for testing.

To evaluate the effectiveness of SMFS on emotion recognition, we conduct experiments on the data generated from three separate sessions of 11 different subjects and compare SMFS with the single-view and multi-view feature selection algorithms. The experiments are repeated 20 times using the same experiment settings as in Section IV-C. Table V presents the average classification accuracy of the RLSC using the top 20 features selected by different feature selection methods. The last column of Table V shows the average performance over all the subjects for each feature selection method. Generally, the semi-supervised multi-view methods mostly outperform the single-view competitors as well as the supervised MSFS, demonstrating the significance of using abundant unlabeled data and complementary information of multiple views for performance improvement. Moreover, SMFS achieves significantly better results compared to all competitors on all subjects for classification accuracy. The single-view version

[2]http://bcmi.sjtu.edu.cn/∼seed/seed.html

TABLE VI
THE LEARNED AVERAGE WEIGHTS OF DIFFERENT BANDS/VIEWS.

Band/View	Delta	Theta	Alpha	Beta	Gamma
MSFS	0.42	0.23	0.20	0.10	0.05
MLSFS	0.20	0.20	0.20	0.20	0.20
MRMVFS	0.33	0.17	0.18	0.17	0.15
MASFS	0.19	0.19	0.19	0.21	0.22
SMFS	0.25	0.12	0.10	0.18	0.35



(a) Gamma Band (b) Delta Band (c) Beta Band

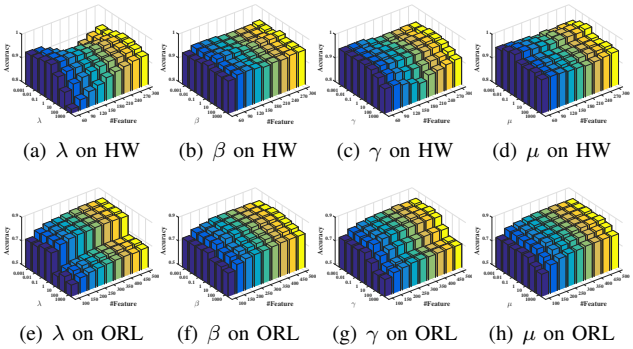
Fig. 4. The top 20 features (marked by red font) selected by SMFS on the Subject 1 are from the Gamma, Delta and Beta bands.

(i.e., SSFS) likewise performs better or comparably than other single-view competitors. Especially, SMFS achieves a 4.6% improvement on Subject 8 in comparison with the best result of all the other competitors, and also achieves 3.6% to 9.2% improvement compared to others in terms of the average performance over all the subjects. The results indicate that the proposed SMFS is more effective in emotion recognition than the state-of-the-art feature selection methods.

According to [51], the neural signals of positive emotions have significantly higher Gamma and Beta responses than those of negative and neutral emotions. Furthermore, the neural signals of neutral emotions have higher Delta responses compared to the negative emotions, indicating that the Gamma, Delta and Beta bands are critical views and more related to these emotions than others. To further evaluate the ability of SMFS for identifying the critical views/bands and features for emotion recognition, we record the learned average weights of different views (i.e., frequency bands) by the multi-view methods on all the subjects in Table VI. From Table VI, we observe that other methods either assign the same weights to different bands or leave out some critical bands. Contrary to this, the proposed SMFS identifies these critical bands and adaptively assigns larger weights to them. This implies that the features from the Gamma, Delta and Beta bands contain more important information for emotion recognition task. Accordingly, Fig. 4 shows the top 20 features selected by SMFS on Subject 1. As depicted in the figure, the selected features are mostly from the Gamma and Delta bands, and their positions are also consistent with the previous studies of critical brain areas [50], [51]. The results indicate that SMFS not only effectively identifies the critical views among multiple views but also the discriminative features within each view.

E. Parameter Sensitivity and Convergence Analysis

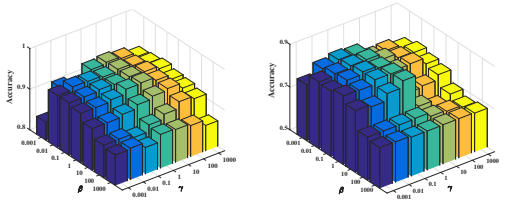
In SMFS, there are four manual parameters λ, β, γ and μ. To investigate the impacts of the parameters and the number of selected features on the performance, we first vary each parameter and the number of selected features. Due to space limitation, the experimental results on the HW and ORL datasets with 10% labeled samples are shown in Fig. 5. It is observable that SMFS has similar performance variation trends on HW and ORL, i.e., ascending initially and descending with the increase in parameters. Moreover, the performance drops significantly with λ greater than 1, indicating the sensitivity of SMFS towards λ. Besides, SMFS is somewhat sensitive to these parameters when the number of selected features is small. To further study the impacts of β and γ on the performance, we fix λ and μ as 1, and set the number of selected features to the median value (i.e., 180 for HW and 300 for ORL). The experimental results are shown in Fig. 6. It can be seen that the performance of SMFS ascends initially and descends with the increase in β and γ. Through tuning the parameters, SMFS achieves relatively good results with β and γ in the range of $\{0.1, 1, 10\}$. When β is larger than 10, the performance has a significant decreasing trend, indicating that the quality of selected features will be degraded if the graph learning is overemphasized. Moreover, SMFS performs worse when the values of β and γ are very small (e.g., 10^{-3}). This demonstrates that the terms corresponding to β and γ are particularly important for identifying discriminative features. Specifically, graph learning can guarantee the accurate propagation of label information on the learned graph, while feature selection aims to learn discriminative projection for fitting prediction labels. Therefore, β and γ should be appropriately set to improve the performance. Considering the different roles of feature selection and graph learning in semi-supervised multi-view feature selection applications, how to automatically determine the optimal β and γ is still an open problem, which will be studied in the future.



(a) λ on HW (b) β on HW (c) γ on HW (d) μ on HW

(e) λ on ORL (f) β on ORL (g) γ on ORL (h) μ on ORL

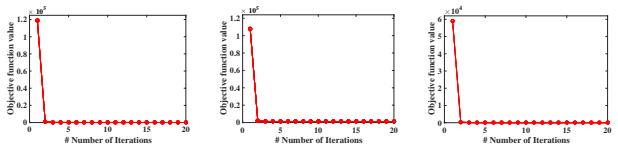
Fig. 5. The accuracy of SMFS with varying parameters and the number of selected features on the HW and ORL datasets with 10% labeled samples.



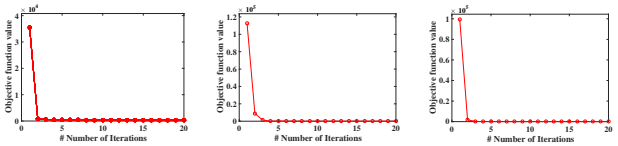
(a) HW (b) ORL

Fig. 6. The accuracy with varying β and γ on the HW and ORL datasets.

Although, we have theoretically proved the convergence of SMFS in Section III-A, here, we further verify its convergence on the benchmark datasets with 10% labeled samples. The



(a) MSRC-v1 (b) HW (c) Cal-7



(d) COIL20 (e) ORL (f) SCENE

Fig. 7. Convergence curves of SMFS with 10% labeled samples.

parameters λ, β, γ and μ are set to 1, and the variation curves of the objective function in Eq. (9) with the number of iterations are shown in Fig. 7. It is clear that the objective function rapidly decreases in the first few iterations and stably converges within 5 iterations on all datasets. The fast convergence guarantees the optimization efficiency of SMFS.

V. Conclusion

In this paper, we propose a novel semi-supervised multi-view feature selection algorithm (SMFS), providing a comprehensive and effective scheme to exploit multi-view data. Different from most existing works which indiscriminately concatenate multiple view features for joint feature selection, and predefine view-specific similarity graphs based on the Euclidean distance between original data points, SMFS not only balances the contributions of different view features, but also effectively coalesces multiple feature projections straightforward. This helps to form a joint feature projection in an adaptive-weighting way by merging the learned weights into the corresponding feature projections, which facilitates preserving the complementarity and consensus among multiple views in the aspect of the feature projections. Furthermore, SMFS adaptively learns a unified similarity graph compatible across multiple views with implicit multiple graph fusion and sample similarity in the joint feature projection space, largely alleviating the adverse effects of irrelevant and redundant features. Moreover, an iterative solution with convergence guaranteed theoretically and demonstrated experimentally is presented to solve SMFS. Extensive experiments on different datasets fully verify that SMFS is effective to select the discriminative features from heterogeneous feature spaces, and achieves superior performance to the state-of-the-arts.

Although SMFS achieves its objectives, some important directions are worth for future research. First, we would like to generalize the $l_{2,1}$-norm to a more feasible $l_{2,p}$-norm constraint where $p \in (0, 1)$ [25]. Second, it is possible to design a more general multi-view feature selection framework that could work in supervised, semi-supervised and unsupervised scenarios. Additionally, the task to extend SMFS to handle large-scale and high-dimensional multi-view data is another important direction in the future.

References

[1] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[2] C. Xu, D. Tao, and C. Xu, "Multi-view intact space learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12, pp. 2531–2544, 2015.

[3] H. Guo, B. Sheng, P. Li, and C. P. Chen, "Multiview high dynamic range image synthesis using fuzzy broad learning system," *IEEE Transactions on Cybernetics*, vol. 51, no. 5, pp. 2735–2747, 2021.

[4] J. Wen, Y. Xu, and H. Liu, "Incomplete multiview spectral clustering with adaptive graph learning," *IEEE Transactions on Cybernetics*, vol. 50, no. 4, pp. 1418–1429, 2020.

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

[6] 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.

[7] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[8] X. Li, Y. Wang, and R. Ruiz, "A survey on sparse learning models for feature selection," *IEEE Transactions on Cybernetics*, vol. 52, no. 3, pp. 1642–1660, 2022.

[9] C. Shi, C. Duan, Z. Gu, Q. Tian, G. An, and R. Zhao, "Semi-supervised feature selection analysis with structured multi-view sparse regularization," *Neurocomputing*, vol. 330, pp. 412–424, 2019.

[10] Y. Li, X. Shi, C. Du, Y. Liu, and Y. Wen, "Manifold regularized multi-view feature selection for social image annotation," *Neurocomputing*, vol. 204, pp. 135–141, 2016.

[11] X. Dong, L. Zhu, X. Song, J. Li, and Z. Cheng, "Adaptive collaborative similarity learning for unsupervised multi-view feature selection," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2018, pp. 2064–2070.

[12] H. Zhang, D. Wu, F. Nie, R. Wang, and X. Li, "Multilevel projections with adaptive neighbor graph for unsupervised multi-view feature selection," *Information Fusion*, vol. 70, pp. 129–140, 2021.

[13] J. Tang, X. Hu, H. Gao, and H. Liu, "Unsupervised feature selection for multi-view data in social media," in *Proceedings of the 2013 SIAM International Conference on Data Mining*, 2013, pp. 270–278.

[14] Y. Wan, S. Sun, and C. Zeng, "Adaptive similarity embedding for unsupervised multi-view feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 10, pp. 3338–3350, 2021.

[15] R. Zhang, F. Nie, X. Li, and X. Wei, "Feature selection with multi-view data: A survey," *Information Fusion*, vol. 50, pp. 158–167, 2019.

[16] H. Tao, C. Hou, F. Nie, J. Zhu, and D. Yi, "Scalable multi-view semi-supervised classification via adaptive regression," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4283–4296, 2017.

[17] F. Nie, L. Tian, R. Wang, and X. Li, "Multiview semi-supervised learning model for image classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 12, pp. 2389–2400, 2020.

[18] X. Xie and S. Sun, "General multi-view semi-supervised least squares support vector machines with multi-manifold regularization," *Information Fusion*, vol. 62, pp. 63–72, 2020.

[19] X. Jia, X. Jing, X. Zhu, S. Chen, B. Du, Z. Cai, Z. He, and D. Yue, "Semi-supervised multi-view deep discriminant representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2496–2509, 2021.

[20] H. Chen, J. Liu, Y. Lv, M. H. Li, M. Liu, and Q. Zheng, "Semi-supervised clue fusion for spammer detection in sina weibo," *Information Fusion*, vol. 44, pp. 22–32, 2018.

[21] J. Zhao, K. Lu, and X. He, "Locality sensitive semi-supervised feature selection," *Neurocomputing*, vol. 71, pp. 1842–1849, 2008.

[22] T. Luo, C. Hou, F. Nie, H. Tao, and D. Yi, "Semi-supervised feature selection via insensitive sparse regression with application to video semantic recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1943–1956, 2018.

[23] B. Jiang, X. Wu, K. Yu, and H. Chen, "Joint semi-supervised feature selection and classification through bayesian approach," in *Proceedings of the 33rd Conference on Artificial Intelligence*, 2019, pp. 3983–3990.

[24] Z. Ma, F. Nie, Y. Yang, J. Uijlings, N. Sebe, and A. Hauptmann, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1662–1672, 2012.

[25] X. Chen, G. Yuan, F. Nie, and Z. Ming, "Semi-supervised feature selection via sparse rescaled linear square regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 1, pp. 165–176, 2020.

[26] C. Shi, Q. Ruan, G. An, and C. Ge, "Semi-supervised sparse feature selection based on multi-view laplacian regularization," *Image and Vision Computing*, vol. 41, pp. 1–10, 2015.

[27] C. Shi, G. An, R. Zhao, Q. Ruan, and Q. Tian, "Multiview hessian semisupervised sparse feature selection for multimedia analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 9, pp. 1947–1961, 2017.

[28] E. Yu, J. Sun, J. Li, X. Chang, X. H. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1276–1288, 2019.

[29] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1747–1756, 2020.

[30] C. Shi, Z. Gu, C. Duan, and Q. Tian, "Multi-view adaptive semi-supervised feature selection with the self-paced learning," *Signal Processing*, vol. 168, p. 107332, 2020.

[31] Z. Kang, H. Pan, S. C. Hoi, and Z. Xu, "Robust graph learning from noisy data," *IEEE Transactions on Cybernetics*, vol. 50, no. 5, pp. 1833–1843, 2020.

[32] K. Zhan, C. Zhang, J. Guan, and J. Wang, "Graph learning for multiview clustering," *IEEE Transactions on Cybernetics*, vol. 48, no. 10, pp. 2887–2895, 2018.

[33] Q. Wang, R. Liu, M. Chen, and X. Li, "Robust rank-constrained sparse learning: A graph-based framework for single view and multiview clustering," *IEEE Transactions on Cybernetics*, 2021.

[34] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Learning a nonnegative sparse graph for linear regression," *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2760–2771, 2015.

[35] Y. Wang, X. Li, and R. Ruiz, "Weighted general group lasso for gene selection in cancer classification," *IEEE Transactions on Cybernetics*, vol. 49, no. 8, pp. 2860–2873, 2019.

[36] M. Luo, X. Chang, L. Nie, Y. Yang *et al.*, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Transactions on Cybernetics*, vol. 48, no. 2, pp. 648–660, 2018.

[37] F. Nie, X. Wang, M. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016, pp. 1969–1976.

[38] F. Dornaika and Y. El Traboulsi, "Joint sparse graph and flexible embedding for graph-based semi-supervised learning," *Neural Networks*, vol. 114, pp. 91–95, 2019.

[39] H. Chen, B. Jiang, and X. Yao, "Semisupervised negative correlation learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5366–5379, 2018.

[40] L. Zhang, M. Luo, Z. Li, F. Nie, H. Zhang, J. Liu, and Q. Zheng, "Large-scale robust semisupervised classification," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 907–917, 2019.

[41] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proceedings of Advances in Neural Information Processing Systems*, 2004, pp. 321–328.

[42] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[43] J. Huang, F. Nie, and H. Huang, "A new simplex sparse learning model to measure data similarity for clustering," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2015, pp. 3569–3575.

[44] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," in *Proceedings of Advances in Neural Information Processing Systems*, 2010, pp. 1813–1821.

[45] B. Jiang, C. Li, M. D. Rijke, X. Yao, and H. Chen, "Probabilistic feature selection and classification vector machine," *ACM Transactions on Knowledge Discovery from Data*, vol. 13, no. 2, pp. 1–27, 2019.

[46] K. Benabdeslem and M. Hindawi, "Efficient semi-supervised feature selection: Constraint, relevance, and redundancy," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1131–1143, 2014.

[47] Y. Zhang, J. Wu, Z. Cai, and S. Y. Philip, "Multi-view multi-label learning with sparse feature selection for image annotation," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2844–2857, 2020.

[48] X. Chang, F. Nie, Y. Yang, and H. Huang, "A convex formulation for semi-supervised multi-label feature selection," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014, pp. 1171–1177.

[49] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A survey on semi-supervised feature selection methods," *Pattern Recognition*, vol. 64, pp. 141–158, 2017.

[50] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2019.

[51] W. Zheng, J. Zhu, and B. Lu, "Identifying stable patterns over time for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 417–429, 2019.

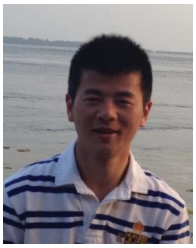
Bingbing Jiang received the BSc degree from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2014, and the PhD degree with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China, in 2019. He is currently working with Hangzhou Normal University, Hangzhou, China. His research interests include semi-supervised learning, multi-view data fusion, sparse learning, feature selection, causal learning and Bayesian inference.

Xingyu Wu received the BSc degree from University of Electronic Science and Technology of China, Chengdu, China, in 2018. He is currently pursuing the Ph.D degree in the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include causal learning, causal inference and feature selection. He has published some scientific papers in prestigious journals and conferences, and served as a Reviewer for IEEE TNNLS and IEEE TKDE, and PC member of AAAI and EMNLP.

Xiren Zhou received the BSc degree from Shandong University, Jinan, China, in 2014, and the PhD degree with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China, in 2019. He is currently an associate researcher with the School of Computer Science and Technology, University of Science and Technology of China. His research interests include machine learning, ground-penetrating radar, and multi-sensor data fusion.

Yi Liu received the MSc degree from Air Force Engineering University, Xi'an, China, in 2011 and the PhD degree from Zhejiang University, Hangzhou, China, in 2020. He is currently working with Hangzhou Normal University, Hangzhou, China. His research interests include data-driven industrial process monitoring and graph model as well as sparse learning.

Anthony G. Cohn received the BSc and PhD degrees in computer science from the University of Essex, Essex, UK. He is a Full Professor with the School of Computing, University of Leeds, Leeds, UK. His research interests include artificial intelligence, knowledge representation and reasoning, cognitive vision, robotics, sensor fusion, and decision support systems. For more than a decade part of his research has focused on decision support systems for street works and utilities. He is a Fellow of the Royal Academy of Engineering, the Association for the Advancement of Artificial Intelligence, and the European Association for Artificial Intelligence. He has received Distinguished Service awards from the International Joint Conferences on Artificial Intelligence and the Association for the Advancement of Artificial Intelligence.



Weiguo Sheng received the MSc degree in information technology from the University of Nottingham, UK, in 2002, and the Ph.D. degree in computer science from Brunel University London, UK, in 2005. He worked as a Researcher at the University of Kent, Canterbury, UK, and Royal Holloway, University of London, London, UK. He is currently a Professor at Hangzhou Normal University, Hangzhou, China. His research interests include evolutionary computation, data mining/clustering, pattern recognition, and machine learning. He is a member of the IEEE.

Huanhuan Chen received the BSc degree from the University of Science and Technology of China, Hefei, China, in 2004 and the PhD degree in computer science from the University of Birmingham, Birmingham, UK, in 2008. He is currently a Full Professor in the School of Computer Science and Technology, USTC. His research interests include neural networks, Bayesian inference and evolutionary computation. Dr. Chen received the 2015 International Neural Network Society Young Investigator Award, the 2012 IEEE Computational Intelligence Society Outstanding Ph.D. Dissertation Award, the 2009 IEEE Transactions on Neural Networks Outstanding Paper Award, and the 2009 British Computer Society Distinguished Dissertations Award. He is an associate editor of the *IEEE Transactions on Emerging Topics in Computational Intelligence*. He is a senior member of the IEEE.