Prior Knowledge Regularized Multiview Self-Representation and Its Applications

Xiaolin Xiao, Yongyong Chen, Yue-Jiao Gong, Member, IEEE, and Yicong Zhou, Senior Member, IEEE

Abstract—To learn the self-representation matrices/tensor that encodes the intrinsic structure of the data, existing multiview self-representation models consider only the multiview features and, thus, impose equal membership preference across samples. However, this is inappropriate in real scenarios since the prior knowledge, e.g., explicit labels, semantic similarities, and weak-domain cues, can provide useful insights into the underlying relationship of samples. Based on this observation, this article proposes a prior knowledge regularized multiview self-representation (P-MVSR) model, in which the prior knowledge, multiview features, and high-order cross-view correlation are jointly considered to obtain an accurate self-representation tensor. The general concept of “prior knowledge” is defined as the complement of multiview features, and the core of P-MVSR is to take advantage of the membership preference, which is derived from the prior knowledge, to purify and refine the discovered membership of the data. Moreover, P-MVSR adopts the same optimization procedure to handle different prior knowledge and, thus, provides a unified framework for weakly supervised clustering and semisupervised classification. Extensive experiments on real-world databases demonstrate the effectiveness of the proposed P-MVSR model.

Index Terms—Low-rank tensor representation, multiview, prior knowledge, self-representation, semisupervised classification, tensor Singular Value Decomposition (t-SVD), weakly supervised clustering.

I. INTRODUCTION

ANY real-world applications are confronted with multiview data as a single-view feature cannot reveal the structure of the data in most cases. For example, in computer vision, multiple heterogeneous features, such as color, texture, and shape, are used to characterize the images. Usually, each view portrays a specific relationship and captures only partial information of the data. Thus, it is necessary to integrate the information from multiple views to explore the underlying relationship of samples [1], [2]. Taking advantage of the complementary information beneath different views, the multiview models have witnessed the performance enhancement over their single-view counterparts in applications of clustering [3]–[15], semisupervised classification [8], [10], [16], [17], and so on.

Among the extensive studies on multiview learning, the self-representation models [3], [5]–[7], [9], [11]–[13], [18] have become the mainstream. Existing models can be characterized by their assumptions on the cross-view correlation and are roughly classified into two categories.

1) Pairwise Correlation Based: For example, Cheng et al. [3] concatenated the representation matrices of different views along the column direction as a large representation matrix and then ensured the sparsity-consistency among all views by imposing the $l_{2,1}$-norm on the concatenated representation matrix. Gao et al. [7] devised an indicator to uncover a common cluster structure agreed by all views. Meanwhile, the models in [5], [9], and [13] emphasized the complementarity of multiview features by exploring the diversity, exclusivity, and nonlinear local manifold structures from different views, respectively.

2) High-Order Correlation Based: To well capture the high-order relationship among samples and across different views, the third-order tensor representation is exploited. Zhang et al. [6] stacked the representation matrices into a third-order tensor and imposed a low-rank constraint on this representation tensor. Relying on the unfolding-based tensor nuclear norm (u-TNN) [19], the low-rank constraint in [6] suffers from the loss of representation optimality [11], [12]. To overcome this limitation, the works in [11] and [12] proposed two optimal strategies to measure the low-rank property of the self-representation tensor.

While effective, previous methods consider only the multiview features and their cross-view relationship. In particular, when calculating the representation coefficients of one sample, existing algorithms assign equal preference for all the other samples. This is inappropriate in the setting of weakly supervised clustering where the membership of samples is constrained via prior knowledge, i.e., weak-domain cues. The domain cues are of great importance in correctly detecting the cluster membership especially when the features are not enough discriminant [3]. Taking region-based image segmentation as an example, images are segmented into superpixels, and the superpixels are clustered into homogeneous segments.
considering multiview features [3]. According to the Gestalt principle of perceptual organization [20], [21], superpixels within one segment should be spatially connected and compact. However, this requirement cannot be satisfied by existing multiview clustering algorithms since they overlooked this valuable prior knowledge. Similar problems may arise in many real-world applications in which the membership preference derived from prior knowledge is informative.

Moreover, in the application of semisupervised classification, it is also critical to associate the available data labels with multiview features. In addition, fine-grained semantic similarities between samples could also provide rich information about the underlying class membership [22]. In these scenarios, the prior information is drawn from explicit data labels or semantic similarities. To exploit the structure information beneath the data labels, Cai et al. [16] proposed a multimodal model to propagate class labels from the labeled samples to the unlabeled ones. By investigating the difficulty of classifying challenging samples, Gong et al. [17] resorted to curriculum learning to boost the performance of multimodal learning. Besides, it is feasible to adapt existing multiview clustering models [4] to the semisupervised scenario [16], [18]. Nevertheless, the adaption always requires tedious works to define new strategies for label propagation, with the exceptions of [8] and [10].

Based on these observations, we propose a novel prior knowledge regularized multiview self-representation (P-MVSR) model that seamlessly integrates the comprehensive information from prior knowledge, multiview features, and the high-order cross-view correlation for multiview learning. As a complement of the multiview features, prior knowledge is used to purify and refine the learned self-representation tensor. In real-world applications, the prior knowledge may come from implicit domain cues, explicit data labels, or semantic similarities, with applications to weakly supervised clustering and semisupervised classification. The novelty and contributions of this article lie in the following aspects.

1) We propose a novel P-MVSR model that takes advantage of the prior knowledge for learning the self-representation tensor. To the best of our knowledge, this is the first model to improve the performance of multiview learning using extra information, which is complementary to the multiview features.

2) P-MVSR employs the same optimization procedure for different prior knowledge and, thus, provides a unified framework for semisupervised classification and weakly supervised clustering, relieving the burden of designing new models for different applications. Besides, the multiview clustering algorithm in [12] falls into our special case when no prior knowledge is imposed.

3) We devise an efficient optimization algorithm to solve P-MVSR via the augmented Lagrangian method with theoretical convergence analysis. Besides, the introduction of prior knowledge will not increase the computation complexity.

4) Within the P-MVSR framework, two illustrative applications are investigated by exploiting the spatial cues in region-based image segmentation and the explicit labels/implicit semantic similarities in semisupervised classification. Extensive experiments have demonstrated the effectiveness and the generalization ability of the proposed P-MVSR framework.

In the rest of this article, Section II reviews the related works on multiview learning. Section III introduces the notations and background knowledge. Section IV elaborates the proposed P-MVSR model. The applications of P-MVSR are demonstrated in Section V. Finally, conclusions are drawn in Section VI.

II. RELATED WORK

Multiview learning [1], [2] aims to learn the intrinsic structure of data from diverse views, from which the consensus and/or the complementary information are well-considered. According to the strategies to integrate multiview features, we classify multiview learning methods into two general groups: 1) multiview representation fusion and 2) multiview representation alignment [2].

The core of multiview representation fusion is to blend the multiview inputs into a single common representation shared by all views. Models belonging to this category differ in the common representation measures and/or fusion schemes. Karasuyama and Mamitsuka [23] proposed to learn a common class indicator matrix by unifying the multiple graph Laplacian matrices with a sparse weighting scheme. To learn the weights of multiple graphs automatically, Nie et al. [8] devised an autoweighting multiple graph learning (AMGL) algorithm to obtain the common class indicator matrix. Afterward, a multiview learning model with adaptive neighbors (MLAN) [10] was introduced to learn the common cluster indicator matrix and the sample similarity matrix simultaneously. Considering each view as one modal, the multimodal learning techniques were employed to find out the hidden pattern from data, e.g., [16] (AMMSS) and [17] (MMCL).

On the other hand, multiview representation alignment captures the relationship among different views via feature alignment. That is, the individual representations of different views are aligned through predefined metrics, such as the correlation measurement [24], the similarity and distance measurements [25]–[27], meaningful customized structures [3], [9], manifold structures [5], [7], [13], [14], [28], and the low-rank tensor constraints [6], [11], [12], [18]. Representatives of multiview alignment are the canonical correlation analysis-based methods [24], [29] and the co-training/co-regularization-based methods [25]–[27]. The abovementioned methods suffer limitations in capturing the high-order cross-view correlation [2]. Inspired by the wide applicability of sparse representation [30], [31] and low-rank representation [32], [33], the self-representation-based methods were extended to the multiview scenarios. Increasing research works were developed in this direction, and a large number of multiview learning algorithms have been proposed based on multiview self-representation [3], [5]–[7], [9], [11]–[13], [18]. According to the assumptions of the cross-view correlation, the multiview self-representation models can be sorted into the pairwise correlation-based methods MLAP [3],
DiMSC [5], ECMSC [9], and the high-order correlation-based ones [6], [11], [12].

To exploit the valuable high-order relationship among the multiview representations, the idea of exploiting low-rank tensor representation for multiview subspace clustering (LT-MSC) was introduced in [6]. However, due to the limitation of the u-TNN, LT-MSC suffers losses in capturing the high-order cross-view correlation. To remedy this situation, Yin et al. [11] devised a new model by directing using the self-expressiveness of the third-order tensor (3rdT-MSC). Alternatively, Xie et al. [12] imposed the tensor singular value decomposition-based tensor nuclear norm (t-SVD-TNN) on the rotated representation tensor to thoroughly explore the high-order cross-view correlation (t-SVD-MSC).

III. NOTATIONS AND PRELIMINARIES

In this section, we first clarify the notations and then briefly review some preliminary works. Throughout this article, we use the calligraphy letters (e.g., \( \mathcal{Z} \)) to represent tensors. A blackboard bold letter \( \mathcal{Z} \) is used as a counterpart of tensor \( \mathcal{Z} \), and they are exploited to represent the original representation tensor and the rotated representation tensor, respectively. The two symbols \( \| \cdot \|_F \) and \( \| \cdot \|_\infty \) denote the u-TNN [19] and the t-SVD-TNN [34], [35], respectively. The uppercase letters (e.g., \( \mathcal{Z} \)) are used to indicate matrices. For the detailed introduction on the tensor algebra, please refer to [12], [35], [49].

A. Low-Rank Tensor Representation-Based Multiview Learning

Traditional multiview learning models [3], [5], [7], [9] can only capture the pairwise correlation of different views. To avoid this limitation, a third-order self-representation tensor can be constructed to simultaneously utilize all views, and its tensor rank is exploited as a metric for multiview representation alignment.

A general formulation of low-rank tensor representation-based multiview learning [?, [6], [11], [12]} is given by

\[
\min_{Z, E} \mathcal{R}(Z) + \lambda \left\| E \right\|_F \quad \text{s.t.} \quad Z \in \Omega, \quad E \in \Xi
\]

(1)

where \( Z \in \mathbb{R}^{n \times n \times V} \) is a third-order tensor constructed by stacking \( V \) representation matrices \( (Z^{(v)}) \) along the third direction; \( Z^{(v)} \in \mathbb{R}^{n \times n} \) measures the self-representation coefficients among \( n \) samples according to the \( v \)th view feature; \( \Omega \) and \( \Xi \) are two compact convex sets; \( \mathcal{R}(Z) \) is used to induce the low-rankness of \( Z \); and \( E \) is the concatenation of error matrices, while \( \| \cdot \|_F \) indicates the regularization strategy, e.g., the squared Frobenius norm \( \| Z \|_F^2 \) can be used to model the Gaussian noise and the \( l_2,1 \)-norm is commonly adopted to deal with sample-specific corruptions and outliers [32]. Within this framework, Zhang et al. [6] exploited the u-TNN \( \mathcal{R}(Z) = \| Z \|_\infty \), which is defined as the sum of the nuclear norms of the matrices unfolded along all modes [19] to capture the high-order cross-view correlation.

B. t-SVD-TNN

While effective in many applications, minimizing u-TNN essentially imposes the low-rank constraint in the matrix SVD-based vector space, resulting in an inadequate representation of the tensor low-rankness [12]. Besides, u-TNN lacks a clear physical meaning. To remedy these problems, the work in [12] exploited the t-SVD-TNN \( \mathcal{R}(Z) = \| Z \|_\infty \) to provide a more accurate low-rank constraint. As shown in Fig. 1, given a tensor \( Z \in \mathbb{R}^{n_1 \times n_2 \times n_3} \), let its t-SVD be \( Z = U \ast S \ast V \), where \( \ast \) is the transpose operator. Then, \( U \in \mathbb{R}^{n_1 \times n_1} \) and \( V \in \mathbb{R}^{n_2 \times n_2} \) are orthogonal tensors, and \( S \in \mathbb{R}^{n_3 \times n_3} \) is an f-diagonal tensor [34], [35].

The t-SVD-TNN is defined as the sum of the singular values of all frontal slices of \( Z_f \)

\[
\| Z \|_\infty = \min_{k=1}^{n_3} \left\| Z_f(:, :, k) \right\|_\infty = \sum_{i=1}^{\min[n_1, n_2]} \sum_{k=1}^{n_3} |S(i, i, k)|
\]

(2)

where \( Z_f = \mathcal{F}(Z, [1, 3]) \) is obtained by applying fast Fourier transformation (FFT) on \( Z \) along the third dimension, and \( Z \) and \( Z_f \) are of the same size.

As reported in [12], it is inappropriate to directly impose t-SVD-TNN on \( Z \) due to the intrinsic circulant algebra underlying t-SVD-TNN. To fix this problem, the dimensionality of \( Z \) is shifted to obtain an \( n \times V \times n \) rotated representation tensor \( \hat{Z} \), as illustrated in Fig. 2.

Remark: The single-view self-representation matrix uncovers the pairwise affinities of samples [32], [36]. By stacking the self-representation matrices from multiple views, the third-order self-representation tensor then captures the high-order relationship among samples and across different views [12]. Conceptually, the t-SVD-TNN imposes a structural constraint on this self-representation tensor to encourage a consensus low-rank tensor structure beneath the third-order affinities of samples. More specifically, the low-rankness of the self-representation tensor can be used to reveal the underlying relationship of samples, somewhat analogous to the low-rank matrix representation [32]. The only difference is that the former employs the high-order relationship, while the latter uses the pairwise relationship.
From this viewpoint, both u-TNN and t-SVD-TNN can be used to explore the high-order cross-view correlation. However, t-SVD-TNN possesses representation optimality due to the following facts.

1) The rank of the tensor is computationally intractable, and t-SVD-TNN has been proven to be the tightest convex relaxation to the $l_1$-norm of the tensor multitank [37, Th. 2.4.1].

2) Adopting the rotation operation, each frontal slice of the rotated representation tensor considers the information among different samples and different views in the Fourier domain. This way, t-SVD-TNN well depicts the complicated relationship between samples and views [12].

Throughout this article, we resort to t-SVD-TNN as a surrogate to replace the rank of the representation tensor for computationally tractability.

IV. PROPOSED P-MVSR

In this section, we introduce our P-MVSR model in which the prior knowledge, for the first time, is used to optimize the representation tensor in the process of multiview learning.

A. Model of P-MVSR

Suppose that $X^{(v)} \in \mathbb{R}^{d_v \times n} (v = 1, 2, \ldots, V)$ is the feature matrix of the $v$th view, $d_v$ denotes the dimension of feature, and $n$ is the number of samples. Our P-MVSR model is presented as

$$
\min_{Z, E} \|Z\|_\oplus + \lambda \|E\|_{2,1}
\quad \text{s.t. } X^{(v)} = \underbrace{X^{(v)} \circ Z^{(v)}}_{\text{Prior}} + E^{(v)}, \quad v = 1, 2, \ldots, V
$$

$$
Z = \phi(\{Z^{(1)}, Z^{(2)}, \ldots, Z^{(V)}\})
$$

$$
E = \underbrace{\{E^{(1)}, E^{(2)}, \ldots, E^{(V)}\}}_{\text{Error}}
$$

(3)

where $\|\cdot\|_\oplus$ is the t-SVD-TNN defined in (2); $P^{(v)}$ is the prior knowledge matrix that describes the membership preference of the $v$th view, and practically, we can use either coarse-grained prior knowledge by assuming $P^{(1)} = \cdots = P^{(V)}$ or fine-grained prior similarity for each specific view; $\circ$ is the Hadamard product operator; $\phi(\cdot)$ stacks the representation matrices $\{Z^{(v)}\}$ into a third-order tensor and then shifts it to obtain an $n \times V \times n$ rotated representation tensor $Z$, as illustrated in Fig. 2; and $E$ is obtained by vertically concatenating the error matrices $\{E^{(v)}\}$. The underlying assumption beneath this concatenation operation is that natural corruptions are always sample-specific, i.e., some data are corrupted, while others are clean [6]. In this viewpoint, we stack $\{E^{(v)}\}$ vertically to enforce the columns of $E^{(1)}, \ldots, E^{(V)}$ to have jointly consistent magnitude values across views. Accordingly, the $l_{2,1}$-norm is used to overwhelm the effect of sample-specific corruptions.

In real-world applications, the prior knowledge may come from any complementary information of the multiview features, e.g., labels, semantic similarities, and domain cues. It provides a meaningful way to assign specific membership preferences across samples when calculating the self-representation coefficients. As a result, explicit or implicit cues will have a direct influence on the discovered relationship of the samples. Within the P-MVSR framework, we can adapt the prior knowledge matrices $\{P^{(v)}\}$ for different applications:

1) For semisupervised classification, the $v$th prior knowledge matrix $P^{(v)}$ is defined as

$$
P^{(v)}_{i,j} = \begin{cases} 
\text{sim}(i, j), & \text{if } i = j \\
0, & \text{if } i \neq j \\
\tau, & \text{otherwise}
\end{cases}
$$

(4)

where $\text{sim}(i, j)$ is the similarity of samples $i$ and $j$; $l_i$ ($i = 1, \ldots, n$) represents the label of the $i$th sample when it is known; and $\tau \in (0, 1)$ is a constant, and it is imposed on the sample pairs when at least one label is unavailable. In practice, different similarity measures will be depicted in Section V-A [see (17) and (18)].

2) In weakly supervised clustering, $\{P^{(v)}\}$ are specified by domain cues; two examples of the domain-specific priors in the application of region-based image segmentation will be introduced in Section V-B [see (19) and (20)].

3) Without prior knowledge, all entries of $\{P^{(v)}\}$ are set to one, and thus, the clustering model in [12] can be considered as a special case of P-MVSR.

B. Solution of P-MVSR

Due to the Hadamard product of $\{P^{(v)}\}$ and $\{Z^{(v)}\}$, it is difficult to solve (3) and is intractable to simultaneously update variables $\{Z^{(v)}\}$ and $E$. We exploit the augmented Lagrange multiplier with an alternating direction minimization scheme [38] to solve (3) for its efficiency and effectiveness. By introducing $V$ auxiliary variables $\{D^{(v)}\}$ and an auxiliary tensor $\mathcal{G}$, the optimization of (3) can be equivalently transformed into the following optimization problem:

$$
\min_{Z^{(v)}, D^{(v)}, E, \mathcal{G}} \|\mathcal{G}\|_\oplus + \lambda \|E\|_{2,1}
\quad \text{s.t. } X^{(v)} = X^{(v)} D^{(v)} + E^{(v)}
$$

$$
D^{(v)} = P^{(v)} \circ Z^{(v)}, \quad v = 1, 2, \ldots, V
$$

$$
Z = \phi(\{Z^{(1)}, Z^{(2)}, \ldots, Z^{(V)}\})
$$

$$
E = \underbrace{\{E^{(1)}, E^{(2)}, \ldots, E^{(V)}\}}_{\text{Error}}
$$

$$
Z = \mathcal{G}.
$$

(5)

The optimal variables $\{Z^{(v)}\}, \{D^{(v)}\}, E,$ and $\mathcal{G}$ can be alternately obtained by minimizing the augmented Lagrange function of (5) as

$$
\mathcal{L}(Z^{(v)}, D^{(v)}, E, \mathcal{G}; \Theta_1, \Theta_2, \Theta_3) = \|\mathcal{G}\|_\oplus + \lambda \|E\|_{2,1}
$$

$$
+ \frac{\rho}{2} \sum_{v=1}^{V} \left\|X^{(v)} - X^{(v)} D^{(v)} - E^{(v)} + \Theta^{(v)} \right\|_F^2
$$

$$
+ \left\|D^{(v)} - P^{(v)} \circ Z^{(v)} + \Theta^{(v)} \right\|_F^2
$$

$$
+ \left\|Z - \mathcal{G} + \Theta_3 \right\|_F^2
$$

(6)
where $\Theta_1$, $\Theta_2$, and $\Theta_3$ are the Lagrange multipliers and $\rho > 0$ is the penalty parameter. More specifically, the optimization of (6) is composed of the following subproblems.

1) $Z^{(v)}$-Subproblem: Fixing other variables except $Z^{(v)}$, the problem reduces to
\[
\min_{Z^{(v)}} \|P^{(v)} \odot Z^{(v)} - B^{(v)}\|_F^2 + \|Z^{(v)} - C^{(v)}\|_F^2 \tag{7}
\]
where $B^{(v)} = D^{(v)} + (\Theta_2^{(v)}/\rho)$ and $C^{(v)} = G^{(v)} - (\Theta_3^{(v)}/\rho)$. Due to the Hadamard product, the optimization procedure is composed of elementwise operations. Without loss of generality, we take the optimization of the $(i, j)$th entry of $Z^{(v)}$ as an example
\[
\min_{Z_{i,j}^{(v)}} (B_{i,j}^{(v)} \ast Z_{i,j}^{(v)} - B_{i,j}^{(v)})^2 + (Z_{i,j}^{(v)} - C_{i,j}^{(v)})^2. \tag{8}
\]
Then, $Z_{i,j}^{(v)} = (P_{i,j}^{(v)} \ast B_{i,j}^{(v)} + C_{i,j}^{(v)})/(1 + P_{i,j}^{(v)})^2$ is obtained by setting derivative of (8) to zero.

2) $D^{(v)}$-Subproblem: Fixing other variables except $D^{(v)}$, (6) reduces to
\[
\min_{D^{(v)}} \|X^{(v)}D^{(v)} - T^{(v)}_1\|_F^2 + \|D^{(v)} - T^{(v)}_2\|_F^2 \tag{9}
\]
where $T^{(v)}_1 = X^{(v)} - E^{(v)} + \Theta_1^{(v)}/\rho$ and $T^{(v)}_2 = P^{(v)} \odot Z^{(v)} - \Theta_2^{(v)}/\rho$. By setting the derivative of (9) to zero, the closed-form solution is given by
\[
D^{(v)*} = (X^{(v)}X^{(v)} + I)^{-1}(X^{(v)}T^{(v)}_1 + T^{(v)}_2). \tag{10}
\]
In practice, we can precalculate $(X^{(v)}X^{(v)} + I)^{-1}$ to avoid extra computation cost.

3) $E$-Subproblem: Fixing other variables except $E$ and concatenating $V$ matrices $\{X^{(v)} - X^{(v)}D^{(v)} + \Theta_1^{(v)}/\rho\}$ along the column direction as a temporary matrix $W, E$ can be obtained by minimizing
\[
\min_{E} \frac{1}{2}\|E - W\|_F^2 + \frac{\lambda}{\rho}\|E\|_{2,1}. \tag{11}
\]
Equation (11) is a group Lasso problem, and we use the Lemma 1 (see [39, Lemma 3.1]) to find the solution.

**Lemma 1:** Given a matrix $W \in \mathbb{R}^{m \times n}$ and a positive scalar $\sigma$, the optimal solution of
\[
\min_{E} \frac{1}{2}\|E - W\|_F^2 + \sigma\|E\|_{2,1}
\]
is obtained at
\[
E^*(\cdot, j) = \begin{cases} \frac{\|W(:, j)\|_2 - \sigma}{\|W(:, j)\|_2}W(:, j), & \text{if } \sigma < \|W(:, j)\|_2 \\ 0, & \text{otherwise.} \end{cases} \tag{13}
\]

4) $G$-Subproblem: Fixing other variables except $G$, the closed-form solution of $G$ can be calculated by optimizing
\[
\min_{G} \frac{1}{\rho}\|G\|_\phi + \frac{1}{2}\|G - F\|_F^2 \tag{14}
\]
where $F = Z + (\Theta_3/\rho)$. Equation (14) is the t-SVD-TNN minimization problem, and it can be solved by using the tensor tubal-shrinkage operator [12], [40] as
\[
G^\ast = C_\phi(F) = U \ast C_\phi(S) \ast V \tag{15}
\]
where $F = U \ast S \ast V$ and $C_\phi(S) = S \ast J$, in which $J \in \mathbb{R}^{n \times n \times V}$ is an f-diagonal tensor whose diagonal element in the Fourier domain is $J(i, i, j) = \max(1 - V/\rho\ast S(i, i, j)), 0)$. 

5) Multipliers and Penalty Parameter: The multipliers and penalty parameter can be updated by
\[
\begin{align*}
\Theta_1^{(v)*} &= \Theta_1^{(v)} + \rho(X^{(v)} - X^{(v)}D^{(v)} - E^{(v)}) \\
\Theta_2^{(v)*} &= \Theta_2^{(v)} + \rho(D^{(v)} - P^{(v)} \odot Z^{(v)}) \\
\Theta_3^{(v)*} &= \Theta_3 + \rho(Z - G) \\
\rho^\ast &= \min\{\beta \ast \rho, \rho_{\max}\}. \tag{16}
\end{align*}
\]

To accelerate the convergence, we employ a continuation scheme [38] to iteratively update the penalty parameter $\rho$ until a maximum value $\rho_{\max}$ is achieved, and $\beta$ is empirically set to 2.

Afterward, the affinity matrix of samples can be formulated as $A = (1/V) \sum_{v=1}^{V}(|Z^{(v)}| + |Z^{(v)}|)$. The whole procedure of P-MVSr is summarized in Algorithm 1.

**Algorithm 1:** P-MVSr

**Input:** multiview feature matrices: $\{X^{(v)}\}$; parameter: $\lambda$; prior knowledge matrices: $\{P^{(v)}\}$; 

**Initialize:** $\{Z^{(v)}\}$, $\{D^{(v)}\}$, $E$, $G$, $\{\Theta_1^{(v)}\}$, $\{\Theta_2^{(v)}\}$, $\Theta_3$ are initialized to 0, $\rho$ is initialized to $10^{-3}$, 

$\rho_{\max} = 10^{10}$, $\beta = 2$, $\epsilon = 10^{-7}$; 

1: while not converged do 
2: for $v = 1$ to $V$ do 
3: Update $Z^{(v)}$ by Eq. (7); 
4: Update $D^{(v)}$ by Eq. (10); 
5: end for 
6: Update $E$ by Eq. (13); 
7: Update $G$ by Eq. (15); 
8: Update $\{\Theta_1^{(v)}\}$, $\{\Theta_2^{(v)}\}$, $\Theta_3$, and $\rho$ by Eq. (16); 
9: Check the convergence conditions 
10: $\max_{v=1}^{V}\|X^{(v)} - X^{(v)}D^{(v)} - E^{(v)}\|_\infty \leq \epsilon$, 
11: $\|Z - G\|_\infty \leq \epsilon$, 
12: $\|Z - G\|_\infty \leq \epsilon$, 
13: end while 

**Output:** affinity matrix $A$.

## C. Discussion

1) Convergence Analysis: The objective function of P-MVSr is coupled with respect to the representation tensor $Z$ since $Z$ is constrained by the t-SVD-TNN, prior knowledge, and the self-expressive property of each view. We introduce $G$ and $\{D^{(v)}\}$ to make $Z$ separable, resulting in four block variables. It has been proven that the direct extension of the augmented Lagrangian multipliers for multiblock convex optimization is not necessarily convergent [41]. Therefore, it is infeasible to strictly prove the convergence properties of P-MVSr. Nevertheless, as verified in pioneer work [6], [12], the convexity of the Lagrange function could guarantee the empirical validity of self-representation-based subspace learning methods to some extent. We will show the empirical convergence curves of P-MVSr in Section V-C2.
2) Computation Complexity: The computation cost of P-MVSR is examined by considering the following subproblems.

1) The $Z^{(v)}$-subproblem for all views costs $O(Vn^2)$ operations.
2) When solving the $D^{(v)}$-subproblem, the inverse of matrix can be calculated in advance, and thus, its cost can be ignored, while the bottleneck lies in the cost of matrix multiplication, which is of order $O(Vn^3)$.
3) The $E$-subproblem costs $O(Vn^2)$ operations per iteration.
4) Considering the $G$-subproblem, it takes $O(Vn^2\log(n))$ operations to calculate 3-D FFT and inverse FFT and $O(V^2n^2)$ operations for SVD; since $n \gg V$ and $\log(n) > V$, $O(V^2n^2)$ is negligible compared with $O(Vn^2\log(n))$.

Therefore, the total computation complexity of P-MVSR is $O(iteVn^3)$, where $ite$ is the number of iterations. Compared with the computation complexity of the method [12], the use of prior knowledge will not lead to extra computation complexity since it only contains elementwise operations.

3) Comparison With Other Low-Rank Representation-Based Multiview Learning Methods: The models in [6] and [12] are most relevant to our work because all of them first stack the multiple self-representation matrices into a third-order tensor to capture the high-order cross-view correlation. However, they differ from motivations, mathematical formulations, and the applications.

1) P-MVSR exploits the prior knowledge that is a complement to multiview features to purify and refine the accuracy of the self-representation tensor, whereas [6] and [12] focus on designing strategies to encode the high-order cross-view correlation.
2) In [6] and [12], the low-rank constraints (with different definitions) are used to optimize the self-representation tensor. By contrast, the self-representation tensor in P-MVSR is jointly constrained by the low-rank constraint and the prior knowledge. Since the prior knowledge can be considered as a complement of the multiview features, the proposed P-MVSR will take advantage of this complementary information for performance enhancement.
3) P-MVSR adopts the same procedure to handle different prior knowledge and provides a unified framework for weakly supervised clustering and semisupervised classification. Meanwhile, [6] and [12] are designed especially for the clustering tasks.

V. EXPERIMENTS

As the proposed P-MVSR incorporates the prior knowledge for learning the self-representation tensor, it is expected to take advantage of the membership preference derived from the prior knowledge to find an accurate class/cluster membership from the data. In this section, two illustrative applications are presented by exploiting different prior knowledge, resulting in the settings of semisupervised classification and weakly supervised clustering, respectively.

A. P-MVSR for Semisupervised Classification

1) Experimental Settings:

a) Setups for P-MVSR: For semisupervised classification, a straightforward way is to exploit data labels as a hard similarity measure. In this setting, (4) results in a coarse-grained prior $P^{(1)}_c = \cdots P^{(v)}_c = \cdots P^{(V)}_c$, and

$$P^{(v)}_c(i,j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \\ \tau, & \text{otherwise} \end{cases}$$

Practically, the semantic similarities between samples could be quite complex to provide rich information about the underlying class relationship. When being applied to multiview hashing, the fine-grained ranking technique has shown promising performance at different semantic levels [22]. Inspired by this, we introduce a fine-grained prior $P^{(v)}_f$ as

$$P^{(v)}_f(i,j) = \begin{cases} \exp\left(-\frac{\|x^{(v)}_i - x^{(v)}_j\|^2}{\sigma^2_f}\right), & \text{if } i = j \\ 0, & \text{if } i \neq j \\ \tau, & \text{otherwise} \end{cases}$$

where $x^{(v)}_i$ denotes the feature vector of the $i$th sample from the $v$th view, and $\sigma_f$ is set to the average of all feature distance pairs. Compared with $P^{(v)}_c$, the core of $P^{(v)}_f$ is to extract membership preference from fine-grained semantic similarities instead of the coarse-grained labels. The selection of other two parameters $\lambda$ and $\tau$ will be examined in Section V-C1. Once the affinity matrix $A$ is obtained using Algorithm 1, we set $-A$ as the distance matrix of samples, and the unlabeled samples are classified using the nearest neighbor classifier.

b) Databases and multiview features: Ten benchmark databases on handwritten digits (HW), scene (MSRC-v1, MITIndoor-67, Scene-15), web images (NUS-WIDE), face (ORL, Yale), animals (AwA), and genetic objects (Caltech101-20, Caltech101) are chosen for experiments. The summary of databases and the corresponding multiview features is reported in Table I. For detail descriptions of the multiview features extracted from these databases, please refer to [10], [12], and [42].

c) Competing algorithms: We compare the performance of $P^{(v)}_c$ and $P^{(v)}_f$ with four state-of-the-arts, i.e., AMMSS [16], AMGL [8], MMCL [17], and MLAN [10]. Among them, AMGL, MLAN, $P^{(v)}_c$, and $P^{(v)}_f$ can be applied to both clustering and semisupervised classification, while AMMSS and MMCL are specialized for the task of semisupervised classification. For a fair comparison, we test the performance of competing algorithms over the recommended parameters on all data sets, respectively. Besides, the ratios of available labeled samples are fixed to the first 10%, 20%, and 30% of all samples, respectively.

2) Classification Accuracy: The best classification results of all algorithms are reported in Tables II and III. It can be observed that $P^{(v)}_c$ and $P^{(v)}_f$ obtain obvious advantages over its competitors. In particular, the performances of AMMSS, MLAN, $P^{(v)}_c$, and $P^{(v)}_f$ are
comparable on HW, and they have much better performance than the other two methods; $P_{c\_MVSR}$ and $P_{f\_MVSR}$ outperform the best peer algorithms by the margins of 5%–10% on MSRV-v1, ORL, and Caltech101-20; and they also achieve more than 10% improvements over the second-best algorithms on the other six databases. Please also note that the improvements of $P_{c\_MVSR}$ and $P_{f\_MVSR}$ are more obvious and clearly significant on the large and complicated databases, e.g., AwA and NUS-WIDE, showing a potential ability to process challenging scenarios. The superiorities of the proposed models own much to the efficient usage of the available data labels in discovering the underlying affinity of the data set.

Comparing the performance of $P_{f\_MVSR}$ to that of $P_{c\_MVSR}$, we can see that using fine-grained semantic similarities can improve the classification accuracy when the data set is small and/or the available data labels are limited. The performance of $P_{f\_MVSR}$ decreases when there are plenty of labeled data. The reason is that with the increase of data labels, directly measuring the fine-grained similarities from view features may not be distinguishable. Several well-designed semantic similarity measures may overcome this limitation. Our future work will investigate this.

### B. P-MVSR for Weakly Supervised Clustering

In terms of weakly supervised clustering, region-based image segmentation is chosen as an application example, in which images are preprocessed to generate superpixels, and the superpixels are clustered into several segments. The reasons to employ this application example lie in two folds: 1) the superpixels of natural images usually lie in a low-rank compared to the images in the other databases; 2) the superpixels are more discriminative in the sense that they capture the semantic information of the images. The superiority of our approach is demonstrated in both classification and clustering tasks.

### TABLE I

**Summary of the Multiview Databases Used for Semisupervised Classification**

<table>
<thead>
<tr>
<th>Content</th>
<th>HW</th>
<th>MSRC-v1</th>
<th>Caltech101-20</th>
<th>ORL</th>
<th>Yale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes</td>
<td>10</td>
<td>7</td>
<td>20</td>
<td>40</td>
<td>15</td>
</tr>
<tr>
<td>Total number</td>
<td>2000</td>
<td>210</td>
<td>2386</td>
<td>400</td>
<td>165</td>
</tr>
<tr>
<td>View 1</td>
<td>Fourier of shape (76 d)</td>
<td>color moment (24 d)</td>
<td>Gabor (48 d)</td>
<td>intensity (4096 d)</td>
<td>400</td>
</tr>
<tr>
<td>View 2</td>
<td>profile correlations (216 d)</td>
<td>GIST (512 d)</td>
<td>wavelet moments (40 d)</td>
<td>LBP (3304 d)</td>
<td>165</td>
</tr>
<tr>
<td>View 3</td>
<td>Karhunen-love (64 d)</td>
<td>CENTRIST (254 d)</td>
<td>-</td>
<td>LBP (3304 d)</td>
<td>-</td>
</tr>
<tr>
<td>View 4</td>
<td>pixel averages (340 d)</td>
<td>LBP (256 d)</td>
<td>HOG (1984 d)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>View 5</td>
<td>Zernike moment (47 d)</td>
<td>GIST (512 d)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>View 6</td>
<td>morphologic (6 d)</td>
<td>LBP (928 d)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AWA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NUS-WIDE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caltech101</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIT1000-67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scene-15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classes</td>
<td>30</td>
<td>25</td>
<td>101</td>
<td>67</td>
<td>15</td>
</tr>
<tr>
<td>Total number</td>
<td>30475</td>
<td>3000</td>
<td>8677</td>
<td>5360</td>
<td>4485</td>
</tr>
</tbody>
</table>

### TABLE II

**Classification Results Over Different Percentages of Labeled Samples on Small Databases**

<table>
<thead>
<tr>
<th>HW</th>
<th>MSRC-v1</th>
<th>Caltech101-20</th>
<th>ORL</th>
<th>Yale</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>20%</td>
<td>30%</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>AMSS</td>
<td>0.9733</td>
<td>0.9757</td>
<td>0.9736</td>
<td>0.8354</td>
</tr>
<tr>
<td>AMGL</td>
<td>0.9056</td>
<td>0.9345</td>
<td>0.9511</td>
<td>0.8039</td>
</tr>
<tr>
<td>XMCL</td>
<td>0.8965</td>
<td>0.9405</td>
<td>0.9705</td>
<td>0.7534</td>
</tr>
<tr>
<td>MLAN</td>
<td>0.9759</td>
<td>0.9788</td>
<td>0.9789</td>
<td>0.8258</td>
</tr>
<tr>
<td>P_{c_MVSR}</td>
<td>0.9783</td>
<td>0.9844</td>
<td>0.9871</td>
<td>0.8783</td>
</tr>
<tr>
<td>P_{f_MVSR}</td>
<td>0.9789</td>
<td>0.9856</td>
<td>0.9871</td>
<td>0.8492</td>
</tr>
</tbody>
</table>

1. Red number indicates the best performance; blue number denotes the second best performance.
2. XMCL run out of memory when being applied to Caltech101-20 on a server with 128 GB memory for constructing $V$ matrices of sizes $d_u \times d_u \times n$ simultaneously: $(48^2 + 40^2 + 254^2 + 1984^2 + 512^2 + 928^2) \times 2386 \times 8 \times (2^{30}) = 91.16$ GB.

### TABLE III

**Classification Results Over Different Percentages of Labeled Samples on Large Databases**

<table>
<thead>
<tr>
<th>AWA</th>
<th>NUS-WIDE</th>
<th>Caltech101</th>
<th>MIT1000-67</th>
<th>Scene-15</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>20%</td>
<td>30%</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>AMSS</td>
<td>0.0422</td>
<td>0.0983</td>
<td>0.1281</td>
<td>0.2047</td>
</tr>
<tr>
<td>AMGL</td>
<td>0.0653</td>
<td>0.0819</td>
<td>0.0985</td>
<td>0.1603</td>
</tr>
<tr>
<td>XMCL</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MLAN</td>
<td>0.0807</td>
<td>0.1019</td>
<td>0.1212</td>
<td>0.2335</td>
</tr>
<tr>
<td>P_{c_MVSR}</td>
<td>0.5251</td>
<td>0.2512</td>
<td>0.25</td>
<td>0.4989</td>
</tr>
<tr>
<td>P_{f_MVSR}</td>
<td>0.2035</td>
<td>0.1924</td>
<td>0.2146</td>
<td>0.5037</td>
</tr>
</tbody>
</table>

MMCL run out of memory when being applied to AWA, Caltech101, MIT1000-67, and Scene-15.
subspace [3], [43] and 2) the domain cues can be easily observed in image segmentation.

1) Experimental Settings:

a) Setups for P-MVSR: Following the Gestalt principle of perceptual organization [21], we adopt two priors to regularize the self-representation tensor within the P-MVSR framework.

1) Adjacent Prior: Superpixels that are adjacent within certain layers have high cluster preference, and we empirically set the number of adjacent layers to 4 in all experiments.

2) Spatial Prior: The cluster preference is calculated according to the spatial distance of two superpixels. These two priors result in the settings of P_a_MVSR and P_s_MVSR, respectively. They are calculated from domain cues rather than specific view features and, thus, can be shared among all views, i.e., \( P^{(i)} = \cdots = P^{(v)} \). Mathematically, the abovementioned two priors are defined as

\[
P_a(i, j) = \begin{cases} 
1, & \text{if } i \text{ and } j \text{ are adjacent} \\
\tau, & \text{otherwise}
\end{cases}
\]

where \( 0 < \tau < 1 \), and

\[
P_s(i, j) = \frac{\exp\left(-\frac{\text{dist}(i, j)}{\sigma^2}\right)}{\tau}, \quad \text{if } i \text{ and } j \text{ are adjacent}
\]

where \( \text{dist}(i, j) \) measures the normalized feature distance between \( i \) and \( j \), \( \sigma \) is fixed to 1 empirically, and \( 0 < \tau < 1 \). Afterward, we set \( P_s(P_s < \tau) = \tau \) to avoid artifacts induced by stratification. Once the affinity matrix \( A \) is formed, we use the spectral clustering algorithm to obtain the final clustering result.

b) Databases and competing algorithms: Four segmentation databases are used in the experiments, namely, Berkeley segmentation data set 500 (BSDS500), the segmentation subset of PASCAL visual object classes 2007 (VOC2007), and Weizmann segmentation data sets with one and two objects (WSD 1obj and WSD 2obj). The corresponding sample numbers are 500, 632, 100, and 100. In addition, eight state-of-the-art multiview clustering algorithms are chosen for comparison, including MLAP [3], DiMSC [5], LT-MSC [6], AMGL [8], ECMSC [9], MLAN [10], 3rdT-MSC [11], and t-SVD-MSC [12].

c) Superpixels and multiview features: Since superpixels generated by different algorithms may capture different visual patterns, we adopt two typical superpixel segmentation methods from different categories, i.e., FH [44] with parameters \([0.8, 200, 100]\) and SLIC [45] with 100 superpixels per image, to show the generalization ability of our P-MVSR model. FH tends to produce irregular superpixels, while SLIC always generates superpixels with similar sizes. To take advantage of multiview features in capturing human perception, three kinds of image features (color, texture, and shape) are extracted, i.e., RGB histogram \((8 \times 8 \times 8 = 512\ d)\), uniform color LBP feature \((177\ d)\) [46], and the Bag-of-Visual-Words (BoW) feature by calculating SIFT [47] at each pixel and then applying the k-means algorithm to generate 100 visual words \((100\ d)\).

2) Quantitative Evaluation: Following the literature [3], four standard evaluation metrics are used for comparison: the probabilistic rand index (PRI), the variation of information (VI), the global consistency error (GCE), and the boundary displacement error (BDE). For PRI, the higher, the better performance; for the other three metrics, the lower, the better.

Considering BSDS500 and VOC2007, since the number of ground-truth segments is unknown, we set the segmentation scale (namely, the number of clusters) to 2, 3, \ldots, 40, respectively. Then, the segmentation results are compared at an optimal data set scale (ODS) for the entire data set and an optimal image scale (OIS), following [48]. The quantitative results of all competing algorithms on BSDS500 and VOC2007 are reported in Table IV. Overall, our P_a_MVSR obtains consistently best performance in most cases, followed by P_s_MVSR. As images in BSDS500 and VOC2007 usually have complex structures and multiple objects, the proposed priors have a natural advantage in obtaining spatially compact segments and, thus, can eliminate the false positive clustering results.

Since images in WSD 1obj and WSD 2obj capture only one and two objects, we fix their cluster numbers as two and three (object segments plus one background segment), respectively. The ODS values are, therefore, the same as the OIS ones, and we only report one result for clarity. Table V presents the quantitative segmentation results of different methods on the BSDS500 and VOC2007 data sets. P_a_MVSR and P_s_MVSR obtain relatively better performance. It can be observed that compared with the results on BSDS500 and VOC2007, the advantage of our P_a_MVSR and P_s_MVSR on WSD is not so significant. This is because images in the BSDS databases contain only one or two object segments, and accordingly, the background segment is likely to spread over all image borders. Thus, the proposed (four-layer) adjacent or spatially compact priors become less effective. By contrast, images in BSDS500 and VOC2007 contain multiple objects, and thus, the priors show superiority. This observation coincides with human perception, and more importantly, it confirms the necessity of an appropriate prior according to specific domain knowledge.

3) Visual Comparison: Examples of the segmented images are shown in Fig. 3, in which superpixels within one segment are labeled with the same color. Specifically, we assign each clustered segment to the label of the most overlapped ground-truth segment. The visual results demonstrate the consistency between the clustered segments and the ground truth. P_a_MVSR and P_s_MVSR obtain better visual results compared with their competitors. More importantly, the positive influence of prior knowledge can be observed. For instance, in P_a_MVSR and P_s_MVSR, the superpixels on the branches of the tree (Image 1), the flowers behind the butterfly (Image 2), and the face and neck of the lady (Image 6) are better grouped compared with peer algorithms. This owns much to the high cluster preference of spatially adjacent/compact superpixels using the proposed priors.

C. Model Analysis

1) Parameter Selection: There are two parameters \( \lambda \) and \( \tau \) that should be tuned for P-MVSR. For semisupervised classification, we empirically set the parameter \( \lambda \) in the
range \([0.01, 0.03, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7]\), and \(r\) is selected from \((0, 1)\) with step \(0.1\), and \(P_{-\text{MVSR}}\) is chosen for analysis. As two illustrative examples, the classification performance of \(P_{-\text{MVSR}}\) over parameters \(\lambda\) and \(\tau\) is reported on the Caltech101-20 and Yale data sets with 10% labeled samples, respectively. We can observe from Fig. 4 that although the parameters \(\lambda\) and \(\tau\) play important roles on the classification accuracy, the results of \(P_{-\text{MVSR}}\) are still stable over local ranges of parameters. In terms of Caltech101-20, the recommended values of \(\lambda\) and \(\tau\) locate within the ranges \([0.01, 0.05]\) and \([0.4, 0.9]\), respectively; for Yale, we suggest to choose \(\lambda\) from \([0.3, 0.7]\) and \(\tau\) from the range \([0.2, 0.8]\). The optimal values of \(\lambda\) on Yale are larger than those on Caltech101-20. This is because the Yale data set contains complicated variations, such as illumination changes and occlusions. By setting a relatively large value of \(\lambda\), our \(P_{-\text{MVSR}}\) model will greatly penalize the error term to achieve better performance. This observation is helpful to tune the parameters of \(P_{-\text{MVSR}}\).

In the application of weakly supervised clustering, \(\lambda\) is empirically selected from \([1.1, 1.3, \ldots, 2.3]\), and \(r\) is chosen from \([0, 0.2, 0.4, 0.8]\) for \(P_{+\text{MVSR}}\) and \(P_{-\text{MVSR}}\). One example of the performance of \(P_{-\text{MVSR}}\) over model parameters on BSDS500 is given in Fig. 5. As can been seen, for image segmentation, the performance of P-MVSR is relatively insensitive to the choice of \(\lambda\) in the range \([1.1, 1.3, \ldots, 2.3]\), while \(\tau\) is recommended to set to 0.2.

2) Empirical Convergence: To examine the convergence of P-MVSR in real scenarios, we define three residuals (i.e., the stopping criterion in lines 10–12 in Algorithm 1)

\[
\begin{align*}
\text{Reconstruction error} & : r_1 = V_{\text{max}} v_{\text{max}} = 1 \{X(v) - \hat{X}(v) D(v) - E(v)\}_\infty \\
\text{Match error 1} & : r_2 = V_{\text{max}} v_{\text{max}} = 1 \{D(v) - P(v) Z(v)\}_\infty \\
\text{Match error 2} & : r_3 = Z - G_\infty.
\end{align*}
\]

Then, we plot the empirical convergence curves of \(P_{+\text{MVSR}}\) on one large database Caltech101 and one small database.

---

### TABLE IV

**Quantitative Segmentation Results on BSDS500 and VOC2007**

<table>
<thead>
<tr>
<th>Method</th>
<th>BSDS500 dataset</th>
<th>VOC2007 dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FID</td>
<td>OAS</td>
</tr>
<tr>
<td>MLAP</td>
<td>0.5658</td>
<td>0.5942</td>
</tr>
<tr>
<td>DmSC</td>
<td>0.5679</td>
<td>0.5187</td>
</tr>
<tr>
<td>LC-MSC</td>
<td>0.5479</td>
<td>0.5069</td>
</tr>
<tr>
<td>AMG</td>
<td>0.5371</td>
<td>0.5007</td>
</tr>
<tr>
<td>ELMSC</td>
<td>0.5272</td>
<td>0.5457</td>
</tr>
<tr>
<td>3s-TMSC</td>
<td>0.5072</td>
<td>0.5272</td>
</tr>
<tr>
<td>s-TMSC</td>
<td>0.4866</td>
<td>0.4566</td>
</tr>
<tr>
<td>P_{+\text{MVSR}}</td>
<td>0.5722</td>
<td>0.5272</td>
</tr>
<tr>
<td>P_{-\text{MVSR}}</td>
<td>0.5571</td>
<td>0.5069</td>
</tr>
</tbody>
</table>

**TABLE V**

**Quantitative Segmentation Results on WSD Databases**

<table>
<thead>
<tr>
<th>Method</th>
<th>WSD 10b dataset</th>
<th>WSD 20b dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FID</td>
<td>OAS</td>
</tr>
<tr>
<td>MLAP</td>
<td>0.5658</td>
<td>0.5942</td>
</tr>
<tr>
<td>DmSC</td>
<td>0.5679</td>
<td>0.5187</td>
</tr>
<tr>
<td>LC-MSC</td>
<td>0.5479</td>
<td>0.5069</td>
</tr>
<tr>
<td>AMG</td>
<td>0.5371</td>
<td>0.5007</td>
</tr>
<tr>
<td>ELMSC</td>
<td>0.5272</td>
<td>0.5457</td>
</tr>
<tr>
<td>3s-TMSC</td>
<td>0.5072</td>
<td>0.5272</td>
</tr>
<tr>
<td>s-TMSC</td>
<td>0.4866</td>
<td>0.4566</td>
</tr>
<tr>
<td>P_{+\text{MVSR}}</td>
<td>0.5722</td>
<td>0.5272</td>
</tr>
<tr>
<td>P_{-\text{MVSR}}</td>
<td>0.5571</td>
<td>0.5069</td>
</tr>
</tbody>
</table>

OAS: optimal scene scale; OAS: optimal image scale.
Fig. 3. Visual comparisons of different algorithms on BSDS500 with five clustered segments per image using FH superpixels.

Fig. 4. Sensitivity analysis of parameters in $P_{c\_MVSR}$ on Caltech101-20 and Yale data sets with 10% labeled samples.

Fig. 5. Sensitivity analysis of parameters in $P_{a\_MVSR}$ on BSDS500.

database Yale as two illustrative examples in Fig. 6. The three residuals yield stable values after 50 iterations, showing that $P_{c\_MVSR}$ converges well in real applications. The average convergence curves of $P_{c\_MVSR}$ and competing algorithms are plotted in Fig. 7. As can be seen, AMMSS and AMGL generally converge within ten iterations, while MLAN and $P_{c\_MVSR}$ obtain stable solutions after 30–50 iterations. MMCL converges within 20 iterations on the Yale database and runs out of memory on the Caltech101 database. It can be observed that $P_{c\_MVSR}$ costs most iterations before convergence. This is because the optimization function of $P_{c\_MVSR}$ has relatively more constraints compared with the competitors.

3) Runtime Comparison: Theoretically, the computation complexity of P-MVSR is $O(ite V n^3)$. That is, the runtime of P-MVSR relates to the number of samples, views, and real iteration numbers. To compare the empirical runtime, all algorithms are implemented using MATLAB and tested on a workstation with Intel Core i9-9920X processor and Ubuntu system. The empirical runtime is reported in Tables VI and VIII, and the theoretical complexity is presented for reference. For semisupervised classification, AMMSS is the most efficient algorithm, and MMCL is the most time-consuming one. The main cost of P-MVSR lies in the operations in the matrix multiplication and the t-SVD-TNN minimization subproblem. To solve computation bottleneck, fast tensor learning methods are needed.
TABLE VI
EMPIRICAL RUNTIME (SECONDS) AND COMPUTATION COMPLEXITY OF DIFFERENT ALGORITHMS ON SEMISUPERVISED CLASSIFICATION

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>HW</th>
<th>MSR-Cv1</th>
<th>Caltech101-20</th>
<th>ORL</th>
<th>Yale</th>
<th>AwA</th>
<th>NUE-WIDE</th>
<th>Caltech101</th>
<th>MIT Indoor-67</th>
<th>Scene-15</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMSS</td>
<td>8.54</td>
<td>0.08</td>
<td>22.96</td>
<td>1.08</td>
<td>0.34</td>
<td>39.67</td>
<td>19.06</td>
<td>277.63</td>
<td>166.56</td>
<td>142.68</td>
<td>O(t \cdot V \cdot n^2)</td>
</tr>
<tr>
<td>AMGL</td>
<td>18.75</td>
<td>0.19</td>
<td>48.62</td>
<td>2.53</td>
<td>0.97</td>
<td>98.67</td>
<td>46.91</td>
<td>557.92</td>
<td>367.92</td>
<td>331.42</td>
<td>O(t \cdot V \cdot n^2)</td>
</tr>
<tr>
<td>MMCL</td>
<td>1125.57</td>
<td>9.32</td>
<td>2672.1</td>
<td>88.95</td>
<td>47.39</td>
<td>33167.3</td>
<td>2129.62</td>
<td>17889.4</td>
<td>12365.8</td>
<td>88695.2</td>
<td>O(t \cdot V \cdot n^2)</td>
</tr>
<tr>
<td>MLAN</td>
<td>47.83</td>
<td>0.39</td>
<td>89.76</td>
<td>4.97</td>
<td>2.21</td>
<td>2012.86</td>
<td>108.76</td>
<td>1234.83</td>
<td>872.75</td>
<td>642.27</td>
<td>O(t \cdot V \cdot n^2)</td>
</tr>
<tr>
<td>P-MVSR</td>
<td>199.82</td>
<td>1.67</td>
<td>451.74</td>
<td>15.22</td>
<td>8.26</td>
<td>6227.68</td>
<td>406.234</td>
<td>4789.65</td>
<td>3783.78</td>
<td>1878.59</td>
<td>O(t \cdot V \cdot n^2)</td>
</tr>
</tbody>
</table>

TABLE VII
ABLATION STUDY ON THE PRIOR KNOWLEDGE AND t-SVD-TNN MODULES FOR SEMISUPERVISED CLASSIFICATION

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>HW</th>
<th>MSR-Cv1</th>
<th>Caltech101-20</th>
<th>ORL</th>
<th>Yale</th>
<th>AwA</th>
<th>NUE-WIDE</th>
<th>Caltech101</th>
<th>MIT Indoor-67</th>
<th>Scene-15</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>no Prior</td>
<td>0.965</td>
<td>0.836</td>
<td>0.832</td>
<td>0.8944</td>
<td>0.8533</td>
<td>0.1726</td>
<td>0.3224</td>
<td>0.3427</td>
<td>0.5972</td>
<td></td>
<td></td>
</tr>
<tr>
<td>concatenation u-TNN</td>
<td>0.9660</td>
<td>0.8413</td>
<td>0.8288</td>
<td>0.9021</td>
<td>0.8056</td>
<td>0.1022</td>
<td>0.3663</td>
<td>0.4127</td>
<td>0.3725</td>
<td>0.5763</td>
<td></td>
</tr>
<tr>
<td>P_MVSR</td>
<td>0.9783</td>
<td>0.8783</td>
<td>0.886</td>
<td>0.9389</td>
<td>0.8867</td>
<td>0.2515</td>
<td>0.4989</td>
<td>0.6859</td>
<td>0.6815</td>
<td>0.7489</td>
<td></td>
</tr>
<tr>
<td>P_MVSR</td>
<td>0.9789</td>
<td>0.8942</td>
<td>0.8902</td>
<td>0.9611</td>
<td>0.94</td>
<td>0.2035</td>
<td>0.5037</td>
<td>0.6372</td>
<td>0.6438</td>
<td>0.7437</td>
<td></td>
</tr>
</tbody>
</table>

For example, Wu et al. [49] proposed an essential tensor learning method for multiview clustering to avoid matrix multiplication. To further reduce the computation cost of low-rank tensor minimization, we can resort to tensor factorization [50] instead of t-SVD-TNN.

When being applied to image segmentation, the average execution time over the whole database is recorded for each algorithm. Since the runtime is closely related to the number of samples, different superpixel segmentation methods will have a direct influence on the empirical runtime. Specifically, FH generates 200–450 superpixels on BSDS200/VOC2007 and around 150–300 superpixels on WSD databases. Meanwhile, SLIC consistently produces 100 superpixels on all databases. As a result, the runtime of competing algorithms varies dramatically when using the FH superpixel, and it is relatively stable when the SLIC superpixel is adopted.

4) Ablation Study:

a) Benefits from prior knowledge and t-SVD-TNN: To start with, we present an ablation study to show the influence of the prior knowledge module and the effectiveness of t-SVD-TNN. Specifically, when no prior knowledge is imposed, the method in [12] is subsumed into the P-MVSR framework; if the t-SVD-TNN module is ablated, we use two schemes to deal with the cross-view relationship: 1) concatenating all view features into a large feature matrix and 2) stacking the representation coefficient matrices and seeking their consensus via u-TNN. These two strategies are denoted by “concatenation” and “u-TNN” in Tables VII and IX. For semisupervised classification, we conduct experiments using 10% labeled samples on all databases. As to image segmentation, the results on the BSDS500 are reported.

b) Multiview features integration: Given features extracted from multiple sources, P-MVSR engages in designing an integration scheme to obtain enhanced performance. Nowadays, the development of deep learning models makes it possible to extract enhanced features by taking advantage of using prior knowledge in learning the affinity matrix.

To further provide an intuitive illustration of the effectiveness of the two modules, we visualize the discovered affinity of samples using every single view and all views without and with prior knowledge, respectively. Following the method in [51], we adopt t-distributed stochastic neighbor embedding (t-SNE) [52] to show the data structure of MSRC-v1 in Fig. 8. We can see that: 1) the underlying class structures cannot be well discovered using only a single view [see Fig. 8(a)–(d)]; 2) integrating the information from multiview features via t-SVD-TNN, the affinity matrices can well preserve the local structures of different classes [see Fig. 8(e) and (f)]; and 3) compared with Fig. 8(e), a more compact and accurate class structure is observed in Fig. 8(f), demonstrating the advantage of using prior knowledge in learning the affinity matrix.

Fig. 7. Average empirical convergence curves of different algorithms on Caltech101 and Yale databases.

Fig. 8. Average empirical convergence curves of different algorithms on Caltech101 and Yale databases.
of the network structure and plenty of samples [53], [54]. A question naturally arises as follows. Will the proposed P-MVSRe model work when features from single views are already well-performed? To answer this question, we extract deep features from three representative networks, i.e., InceptionV3 (mixed10 layer), ResNet-50 (activation-49 layer), and VGG19 (fc2 layer). All deep models are trained on ImageNet.

Since these models have different structures, we consider the extracted features as three heterogeneous features. We use the Caltech101-20 database as an illustrative example and record the results from the best-performing single view, P-MVSRe, and their performance gains, respectively, in Table X. Although a single deep feature shows promising performance, considerable improvements are witnessed by integrating the features from multiple deep models. This validates the general assumption on integrating multiview features for performance enhancement.

**VI. CONCLUSION**

This article proposed a P-MVSRe model to take advantage of the prior knowledge in discovering the underlying relationship of samples. Specifically, we considered the prior knowledge, multiview features, and the high-order cross-view correlation simultaneously to learn an accurate self-representation tensor. Taking the valuable prior information as a complement to the multiview features, P-MVSRe has shown the superiority compared with existing multiview learning models. Extensive experiments on semisupervised classification and region-based image segmentation have demonstrated the effectiveness and the generalization ability of our P-MVSRe model by investigating different settings of prior knowledge.

As P-MVSRe generalizes different kinds of side information, e.g., explicit labels, semantic similarities, weak-domain cues, as the prior knowledge, and uses the prior knowledge to guide the learning procedure, it opens up new opportunities of developing powerful multiview self-representation models. In the future, we can investigate the application of P-MVSRe in constrained clustering where the task-dependent constraints provide valuable information on the relationship of data. In addition, inspired by the work in [55]–[57], it is worth exploring the fine-grained prior knowledge from the initial coarse priors for performance enhancement.

**REFERENCES**


---

**TABLE IX**

<table>
<thead>
<tr>
<th></th>
<th>PRI</th>
<th>Vol</th>
<th>GCS</th>
<th>BDS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ODS</td>
<td>OIS</td>
<td>ODS</td>
<td>OIS</td>
</tr>
<tr>
<td>no Prior</td>
<td>0.7363</td>
<td>0.784</td>
<td>1.9386</td>
<td>2.1868</td>
</tr>
<tr>
<td>concatenation</td>
<td>0.7375</td>
<td>0.7902</td>
<td>1.7922</td>
<td>2.2089</td>
</tr>
<tr>
<td>u-TNN</td>
<td>0.7465</td>
<td>0.7951</td>
<td>1.9652</td>
<td>2.2012</td>
</tr>
<tr>
<td>P*-MVSR</td>
<td>0.7536</td>
<td>0.8131</td>
<td>1.8939</td>
<td>2.1922</td>
</tr>
<tr>
<td>P+ MVSR</td>
<td>0.7544</td>
<td>0.8045</td>
<td>1.9496</td>
<td>2.248</td>
</tr>
</tbody>
</table>

**TABLE X**

<table>
<thead>
<tr>
<th></th>
<th>Hand-crafted feature</th>
<th>Deep feature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Best-view</td>
<td>P-MVSRe</td>
</tr>
<tr>
<td>10%</td>
<td>0.6769</td>
<td>0.886</td>
</tr>
<tr>
<td>20%</td>
<td>0.7611</td>
<td>0.9291</td>
</tr>
<tr>
<td>30%</td>
<td>0.7974</td>
<td>0.9321</td>
</tr>
</tbody>
</table>

---

1https://keras.io/zh/applications/

2By extending the single-view self-representation model in [32] to the semisupervised scenario.


Xiaolin Xiao received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree from the University of Macau, Macau, China, in 2019. She is currently a Post-Doctoral Fellow with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Her research interests include multiview learning and color image processing and understanding.
Yongyong Chen received the B.S. and M.S. degrees from the College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, China. He is currently pursuing the Ph.D. degree with the Department of Computer and Information Science, University of Macau, Macau, China.

He visited the National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China, as an Exchange Student, in 2017. His research interests include (nonconvex) low-rank and sparse matrix/tensor decomposition models, with applications to image processing, data mining, and computer vision.

Yue-Jiao Gong (Member, IEEE) received the B.S. and Ph.D. degrees in computer science from Sun Yat-sen University, Guangzhou, China, in 2010 and 2014, respectively.

She is currently a Full Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou. Her research interests include evolutionary computation, swarm intelligence, and their applications to intelligent transportation and smart city scheduling. She has published over 80 articles, including more than the 30 IEEE TRANSACTIONS, in these areas.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees from Tufts University, Medford, MA, USA, all in electrical engineering.

He is currently an Associate Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a Senior Member of the International Society for Optical Engineering (SPIE). He was a recipient of the Third Price of Macau Natural Science Award in 2014. He is also the Co-Chair of the Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He also serves as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and four other journals.