IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received August 13, 2019, accepted September 5, 2019, date of publication September 13, 2019,
date of current version September 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941219

Robust Semi-Supervised Non-Negative Matrix Factorization With Structured Normalization

LIUJING WANG[1], NAIYANG GUAN[2,3], (Member, IEEE), DIANXI SHI[2,3], ZUNLIN FAN[2,3], AND LONGFEI SU[2,3]

[1]School of Computer Science, National University of Defense Technology, Changsha 410073, China
[2]Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, Beijing 100073, China
[3]Tianjin Artificial Intelligence Innovation Center, Tianjin 300457, China

Corresponding authors: Naiyang Guan (nyguan@sina.com) and Dianxi Shi (dxshi@nudt.edu.cn)

This work was supported by the National Key Research and Development Program of China through the Unmanaged Cluster System Software Architecture and Operational Support Platform Project under Grant 2017YFB1001901 and in part by the The National Natural Science Foundation of China under the Grant 61502515.

ABSTRACT Non-negative matrix factorization (NMF) approximates a non-negative data matrix with the product of two low-rank non-negative matrices by minimizing the cost of such approximation. However, traditional NMF models cannot be generalized in the cases when the dataset contains outliers and limited knowledge from domain experts. In this paper, we propose a robust semi-supervised NMF model (RSS-NMF) to overcome the aforementioned deficiency. RSS-NMF utilizes the L_2/L_1-norm to encourage approximation and makes the model insensitive to outliers by prohibiting them from dominating the cost function. To incorporate the discriminative information, RSS-NMF utilizes the structured normalization method when learns a diagonal matrix to normalize the coefficients such that they get close to the label indicators of the given labeled examples. Although the multiplicative update rule (MUR) can be adopted to minimize RSS-NMF, it converges slowly. In this paper, we adopt a fast gradient descent algorithm (FGD) to optimize RSS-NMF and prove its convergence to a stationary point. FGD uses a Newton method to search the optimal step length and thus, FGD converges faster than MUR. The experimental results show the promise of RSS-NMF comparing with the representative clustering models on several face image datasets.

INDEX TERMS Non-negative matrix factorization, semi-supervised learning, L_2/L_1-norm, structured normalization.

I. INTRODUCTION

Semi-supervised clustering is a longstanding problem in machine learning filed and has extremely widespread applications, ranging from document processing [36], segmentation [37], behavioral analysis [38], to face recognition [39]. It is a learning method which fully utilizes the prior knowledge to clustering procedure. Based on semi-supervised learning and clustering analysis, the clustering performance can be improved. The semi-supervised algorithms are generally divided into two categories: (1) the constraint-based algorithms, which utilize constraint information to optimize clustering effect, such as SemiSync [31] and MSAEClust [30]; (2) the distance-based algorithms, which learn a distance measure by the assistance of prior kowledge, such as SCKMM [40] and K-EDML [41].

Recently, non-negative matrix factorization(NMF, [1]) has been applied to semi-supervised clustering consistently and has been demonstrated excellent clustering performance. The non-negativity constraint incorporated on two factor matrices induces parts-based representation, which is consistent with the intuition of learning parts to form a whole in human brain [7]–[9]. Since standard NMF minimizes the squared L_2-norm of the approximation error between data matrix and the product of two factor matrices, it is sensitive to outliers as in this case the contaminated entries dominate the objective function.

Over the past decades, many models have been proposed to improve the robustness of NMF. Kong *et al.* [3] proposed an $L_{2,1}$-norm based NMF model ($L_{2,1}$-NMF) which replaces the L_2-norm in classical NMF model with $L_{2,1}$-norm. Since $L_{2,1}$-NMF measures the loss by summing the L_2-norm of columns of the error matrix, it prohibits outliers from dominating the objective function. Lam [19] proposed an L_1-norm based NMF model (L_1-NMF) which measures the loss by

The associate editor coordinating the review of this manuscript and approving it for publication was Ananya Sen Gupta.

using the L_1-norm of the error matrix. Since L_1-NMF avoids the domination of outliers in the objective function, it is more robust than classical NMF models. Hamza and Brady [6] proposed a robust NMF model (RNMF) which measures the loss by summing the L_2/L_1-norm of errors. Since the L_2/L_1-norm gets close to the squared L_2-norm when error goes to zero and gets close to the L_1-norm when error goes to infinity, RNMF inherits the advantages of both classical NMF and L_1-NMF.

The above NMF variants are intrinsically unsupervised models, that is, they do not make use of any discriminative information to promote the learning process. In recent years, many studies [2], [11], [12] have shown that the learning quality of NMF can be enhanced significantly by using a small amount of labeled data. In this situation, many semi-supervised NMF methods are proposed. CNMF [12] regards the label information as additional constraint, namely, samples from the same class are supposed to be mapped to the same representation in the new data representation. However, CNMF fails in case that the label information is rather limited. SCNMF [42] is proposed to solve the above problem. It softens the hard constraint in CNMF by introducing a diagonal matrix with positive diagonal elements to normalize the decomposition. These semi-supervised NMF models incorporate prior information only, but are not robust to noises and outliers.

In this paper, we propose a robust semi-supervised NMF (RSS-NMF), and the clustering process of which is displayed in Figure 1. This model utilizes the L_2/L_1-norm [6] to measure the cost of NMF approximation and make the model insensitive to outliers by prohibiting them from dominating the cost function. Furthermore, RSS-NMF aims at improving the clustering performance by incorporating discriminative information. More concretely, structured normalization regularization is provided to learn a diagonal matrix to normalize the coefficients such that they get close to the label indicators of the given labeled examples. RSS-NMF jointly learns the representations from both labeled examples and unlabeled examples in the presence of outliers on both sides.

Many works have applied the multiplicative update rule algorithm (MUR) to optimize the robust NMF models. Kong *et al.* [3] proposed a power method based MUR to optimize $L_{2,1}$-NMF which updates each factor matrix by MUR with an adaptive re-weighting strategy. Du *et al.* [10] applied a half-quadratic based MUR for optimizing several robust NMF models including correntropy induced metric based NMF, Huber function based NMF, Welsh function based NMF, and Cauchy function based NMF. Although MUR decreases the corresponding objective function, they do not guarantee convergence. For optimizing L_1-NMF, Lam [5] reformulated the objective function and optimized this model by using linear programming (LP). Although the LP algorithm implicitly converges to a local minimum, [5] does not give explicit proof. For optimizing RNMF, Hamza and Brady [6] applied the gradient descent algorithm with a smartly chosen step size. However, the convergence is not guaranteed. For robust NMF model, the convergence of optimization algorithm is important because convergent algorithm makes the learned factor matrices less influenced by initialization.

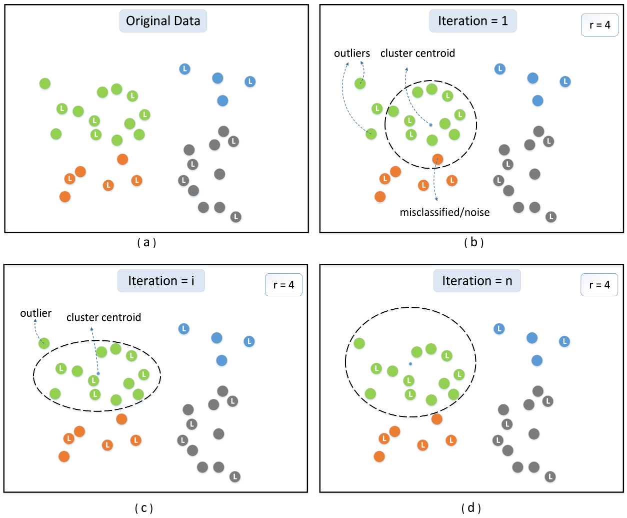


FIGURE 1. Clustering process via RSS-NMF with the noise interference. Suppose the data points are divided into four categories, which are distinguished with each other by four colors. The circles containing a letter 'L' represent the labeled examples, and the outliers are marked with text. (a) Original data points. (b) Clustering result after the first iteration. (c) Clustering result after the i-th iteration. (d) The final clustering result.

To optimize RSS-NMF, we present a fast gradient descent algorithm (FGD) for much faster convergence. Since the objective function of RSS-NMF is non-convex, FGD alternatively updates one factor matrix with another one fixed. For updating each factor matrix, FGD first constructs an auxiliary matrix to make the newly updated matrix satisfy K.K.T. conditions, and then searches along the scaled negative gradient with a suitable step size. The optimal step size is determined by line search based on the Newton algorithm. For theoretically analyzing the convergence of RSS-NMF, we first prove that FGD decreases the objective function, and then prove that any limit points of the generated sequence are stationary points. By showing that the generated sequence has at least one limit point, we prove that FGD converges to a stationary point. Experimental results on both synthetic and real-life datasets demonstrate that FGD is efficient.

This paper is organized as follows: Section II briefly reviews the related works; Section III proposes the robust semi-supervised NMF model (RSS-NMF) and proposes a multiplicative update rule algorithm (MUR) for optimizing RSS-NMF, and then proposes a fast gradient decent (FGD) to accelerate MUR; Section IV empirically evaluates RSS-NMF by showing its efficiency and effectiveness; Section V concludes this paper.

II. RELATED WORKS

In this section, we review clustering, semi-supervised clustering, semi-supervised NMF and robust NMF variants.

We theoretically analyze their loss functions for understanding their advantages as well as disadvantages.

A. CLUSTERING AND SEMI-SUPERVISED CLUSTERING VARIANTS

Clustering is an unsupervised learning problem which divides examples into several groups according to intrinsic characteristics or similarity. Recently, several clustering algorithms have been proposed, such as partitional clustering [46], [47], density-based clustering [23], [24], and clustering based on non-negative matrix factorization [48], [26]. Inspired by deep clustering, Guo *et al.* [32] proposed an improved deep embedded clustering (IDEC) model, which maintains feature space by using a under-complete autoencoder and a clustering loss as guidance.

In the past few years, extensive studies have shown that once a small amount of prior knowledge about the data is incorporated, the performance of clustering can be improved greatly [28], [43]. In order to use the background knowledge of data points reasonably, semi-supervised clustering is proposed. In general, there are generally two types of semi-supervised clustering methods:

1) CONSTRAINT-BASED SEMI-SUPERVISED CLUSTERING

Such methods add constraint restriction information to clustering procedure. The supervision information is divided into two major categories.

The first category is independent class label. The original data set V is expressed as $V = \left[V^l, V^u\right] \in \Re^{m\times n}$ with $n = l + u$, where V^l represents the labeled samples set and V^u represents the unlabeled samples set. Usually, $l \ll u$, the number of labeled samples is much smaller than the number of unlabeled samples. Inspired by the above idea, MSAEClust [30] was proposed, which directly exploits background knowledge. It trains multiple autoencoders of different sizes to incorporate samples together with label information. SSC-SR [33] was presented as a constrained optimization model and was solved via the inexact augmented Lagrangian multiplier (IALM). With the guidance of a small amount of supervision information, SSC-SR utilizes a matrix with anti-block-diagonal appearance to regularize the product of the low-dimensional embedding and its transpose.

The second category is pairwise constraints, which are formalized as instance-level Must-Link constraints(ML) and Cannot-Link(CL). ML indicates that two data points belongs to the same group, while CL is the opposite. That is, given two data points v_i and x_j belongs to class K_i and K_j respectively, if $\left(v_i, v_j\right) \in$ Must-Link, $i = j$, and if $\left(v_i, v_j\right) \in$ Cannot-Link, $i \neq j$. Zhang *et al.* [31] proposed a novel model, namely SemiSync, which focuses on an interesting phenomenon synchronization. SemiSync combines Cannot-Link and Must-Link constraints by introducing a global interaction paradigm. Constraints are propagated within the synchronized in a reasonable way and therefore SemiSync achieves high-quality clustering with limited knowledge is available.

2) DISTANCE-BASED SEMI-SUPERVISED CLUSTERING

Distance-based semi-supervised clustering methods exploit a specific distance metric to satisfy the prior knowledge and then utilizes an existing clustering algorithm to learn the similarity between data points. LSSC [49] develops a new metric function by using pairwise constraints and labeled data. Yin *et al.* [45] develop a semi-supervised fuzzy clustering algorithm with metric learning and entropy regularization simultaneously (SMUC). Yan *et al.* [44] presented a similarity metric clustering method, which utilizes various viewpoints.

B. ROBUST NMF VARIANTS

To remedy the non-robustness of NMF, Kong *et al.* [3] proposed a $L_{2,1}$-norm based NMF ($L_{2,1}$-NMF) which takes off the squared operator. The objective function of $L_{2,1}$-NMF is

$$\min_{W\geq 0, H\geq 0} \sum_{j=1}^{n} \left\|(V-WH)_{\cdot j}\right\|_2^2 . \tag{1}$$

Although $L_{2,1}$-NMF is more robust than NMF, it still cannot filter out intra-sample outliers. For example, when one pixel in an image is seriously corrupted, although this image does not destroy the whole model, the corrupted pixel will conceal the effects of the rest pixels.

Lam [5] proposed a L_1-norm based NMF (L_1-NMF) which minimizes the L_1-norm of the residual error of each sample, i.e.,

$$\min_{W\geq 0, H\geq 0} \sum_{j=1}^{n} \left\|(V-WH)_{\cdot j}\right\|_1 , \tag{2}$$

where $\left\|X_{\cdot j}\right\|_1 = \sum_{i=1}^{m} \left|X_{ij}\right|$ for any j. In contrast to $L_{2,1}$-NMF, L_1-NMF successfully inhibits the influences of both outlier samples and intra-sample outliers. The main shortcoming of L_1-NMF is that it is difficult to be optimized because the absolute function is non-differentiable at zero.

Hamza and Brady [6] proposed a L_2/L_1-norm based NMF model (L_2/L_1-NMF) which replaces the absolute function in (2) with the differentiable L_2/L_1-norm, i.e.,

$$\min_{W\geq 0, H\geq 0} \sum_{j=1}^{n} \sum_{i=1}^{m} \rho\left((V-WH)_{ij}\right), \tag{3}$$

where

$$\rho(x) = \sqrt{1+x^2} - 1 \tag{4}$$

is the Hypersurface cost function. Comparing with the above models, RNMF has three advantages: (1) it is more robust to outlier-samples than NMF because it prohibits large errors to dominate the objective function; (2) it is more robust to intra-sample outliers than $L_{2,1}$-NMF because it prohibits large error entries to dominate the objective function; (3) it is differentiable everywhere and thus can be easily optimized by using gradient based algorithms. L_2/L_1-NMF performs more robustly than NMF. However, the gradient descent algorithm utilized in [6] is complex because the Armijo rule based line search is complex.

C. SEMI-SUPERVISED NMF VARIANTS

The aforementioned basic NMF and robust NMF aim at finding parts-based and linear representations of non-negative data, which take no account of the label information of samples. Real world data are often sparse and noisy, which may reduces the accuracy of data representations. And a small part of data may have prior label information, which, if utilized, may improves the discriminability of representations.

SSNMF [11] incorporated the data matrix and the partial class label matrix into NMF. Associated labels are encoded in the label matrix Y $= [y_1, \ldots, y_n]$, where each y_i is a binary vector such that only the j-th entry is one and remaining elements are zero if x_i belongs to class j. They consider a joint factorization of the data matrix V and the label matrix Y, sharing a common factor matrix H. The loss function of SSNMF is

$$\mathrm{L(W, H, U)} = \|V - WH\|^2 + \lambda\|Y - UH\|^2, \tag{5}$$

where λ is a tradeoff parameter determining the importance of the supervised term.

Liu and Wu [12] introduced constrained nonnegative matrix factorization (CNMF), which merges the label information as additional constraints. CNMF considers a dataset consisting of n data points, among which the label information is available for the first l data points $v_1, \ldots, v_l$, and the rest of the n-l data points $v_{l+1}, \ldots, v_n$ are unlabeled. The data points v_i and v_j have the same low-dimensional representation $h_i = h_j$ if they belong to the same class. Specifically, C is an $l \times c$ indicator matrix where $c_{i,j} = 1$ if x_i is labeled with the j-th class and $c_{i,j} = 0$ otherwise. H is first separated into two parts: $H_{1:m}$ (labeled) and $H_{m+1:n}$ (unlabeled). CNMF requires that $H_{1:m} = QC$ for some nonnegative matrix Q. The reconstruction coefficients $H_{m+1:n}$ for unlabeled examples are not constrained except to be generally nonnegative. Both these conditions can be expressed by

$$\mathrm{H = PA} \quad \text{where } \mathrm{P} = \begin{pmatrix} Q & H_{m_1:n} \end{pmatrix} \text{ and } A = \begin{pmatrix} C & 0 \\ 0 & I_{n-l} \end{pmatrix}, \tag{6}$$

where I_{n-l} is an $(n-l)\times(n-l)$ identity matrix. By plugging H in eq. (6) into the sum of squared errors of the original NMF, the loss function is

$$\mathrm{L(V, P)} = \|V - WPA\|, \tag{7}$$

which is minimized by simple multiplicative updates for W and P.

Chen *et al.* [13] proposed a semi-supervised approach for clustering based on non-negative matrix factorization, which incorporated the pairwise constraints into the similarity matrix of the data. Users are able to provide supervision for clustering in terms of pairwise constraints on a few data objects specifying whether they must or cannot be clustered together. SEMINMF [14] as another variation of CNMF, not only utilizes the local structure of the data characterized by the graph Laplacian, but also incorporates the label information as the fitting constraints to learn.

Symmetric NMF (SNMF) had shown to be effective for graph representation. Wu *et al.* [2] proposed a novel SNMF-based semi-supervised clustering method, named PCPSNMF. PCPSNMF incorporates a small amount of supervisory information into the learned subspace to guide the construction of the similarity matrix. And then the two matrices communicate with each other to achieve mutual refinement until convergence.

Li *et al.* [4] proposed a robust structured NMF learning framework, which learns a robust discriminative representation by leveraging the block-diagonal structure and the $L_{2,p}$-norm loss function. The $L_{2,p}$-norm loss function solve the problems of noise and outliers effectively.

III. ROBUST SEMI-SUPERVISED NMF

In this section, we introduce the proposed RSS-NMF, which not only utilizes the local structure of the data, but also encodes discriminative information of different clusters. Then we apply multiplicative update rule (MUR) method to minimizing RSS-NMF. Finally, a fast gradient descent (FGD) is proposed to accelerate MUR.

A. STRUCTURED NORMALIZATION

Given l labeled examples concatenated in a non-negative matrix $V^l \in \Re^{m\times l}$ and u unlabeled examples concatenated in a non-negative matrix $V^u \in \Re^{m\times u}$, RSS-NMF concatenates them in single non-negative matrix $V = \left[V^l, V^u\right] \in \Re^{m\times n}$ with $n = l + u$, and factorizes V into the product of a basis matrix $W \in \Re^{m\times r}$ and a coefficient matrix $H \in \Re^{r\times n}$. According to the composition of V, the coefficient matrix H is divided into two parts, i.e., $H = \left[H^l, H^u\right]$, where $H^l \in \Re^{r\times l}$ and $H^u \in \Re^{r\times u}$.

For the labeled examples, RSS-NMF encodes their discriminative information in a class indicator matrix as follows:

$$Y^l_{kj} = \begin{cases} 1, & \text{if class } (j) = k \\ 0, & \text{if class } (j) \neq k, \end{cases} \quad \forall 1 \leq k \leq r,\ \forall 1 \leq j \leq l. \tag{8}$$

where class$(x) = k$ means that example x belongs to the class k.

To incorporate the discriminative information of the labeled examples, we expect that the coefficient matrix of labeled examples equals the class indicator matrix, i.e.,

$$H^l = Y^l. \tag{9}$$

However, the above hard constraint is too strict and may makes the learned basis shrink to the labeled examples. Intuitively, in an extreme case, assuming each class contains one labeled example, i.e., $l = r$, and the k-th labeled example belongs to the k-th class, the factorization corresponding to the labeled examples V^l is formulated as follows:

$$V^l \approx WH^l. \tag{10}$$

According to (9) and (10), Y^l is an identical matrix, and thus W shrinks to V^l. To reduce the risk of such shrinkage, we relax the hard constraint (9) by introducing a positive normalizer in learning the coefficient matrix H^l, i.e.,

$$\min_{\Lambda} \frac{1}{2}\left\|\Lambda H^l - Y^l\right\|_F^2, \tag{11}$$

where Λ is a positive diagonal matrix, which normalizes both W and H such that the discriminative information of the labeled examples is incorporated into NMF. Therefore, we call the problem (11) structured normalization. With (11), we get the objective function of RSS-NMF as follows:

$$F(W,H)=\sum_{i=1}^{m}\rho\left(\left(V-W\Lambda^{-1}\Lambda H\right)_{ij}\right)+\frac{1}{2}\left\|\Lambda H^l-Y^l\right\|_F^2. \tag{12}$$

Theorem 1 shows that the loss function of L_2/L_1-NMF is close to that of L_1-NMF, and their distance is bounded. This analysis implies that L_2/L_1-NMF behaves like L_1-NMF, and is therefore more robust to outlier-samples and intra-sample outliers than NMF.

Theorem 1: For any $x \in \Re^m$, we have: $\|x\|_1 - m \leq \sum_{i=1}^{m} \rho(x_i) \leq \|x\|_1$.

It is obvious that (12) is non-convex with respect to W, H, and Λ. Therefore, we apply the block coordinate descent to alternating update each variable with other variables fixed. Since the normalizer Λ has no relationship with W and H, it can be updated separately by

$$\Lambda_{kk} = \frac{\left\langle H^l_{k\cdot}, Y^l_{k\cdot}\right\rangle}{\left\langle H^l_{k\cdot}, H^l_{k\cdot}\right\rangle}, \quad \forall 1 \leq k \leq r. \tag{13}$$

With the normalizer Λ fixed, the factors W and H can be updated by solving the L_2/L_1-NMF model (12). Although the projected gradient descent algorithm (PGD) with Armijo rule has been applied to solve L_2/L_1-NMF, its computational complexity is too high because the Armijo rule needs to compute the objective value in each attempt of searching a suitable step size. In this section, we first propose a multiplicative update rule algorithm (MUR) to solve RSS-NMF. In the next section, we will propose a fast gradient descent algorithm (FGD) to efficiently solve RSS-NMF and prove its convergence to a stationary point.

B. MULTIPLICATIVE UPDATE RULE FOR OPTIMIZING W AND H

Let $F(W,H)=\sum_{j=1}^{n}\sum_{i=1}^{m}\rho\left((V-WH)_{ij}\right)+\frac{1}{2}\left\|\Lambda H^l-Y^l\right\|_F^2$ denote the objective function of (12). Since $F(W,H)$ is jointly non-convex with respect to W and H, we solve (12) by alternatively updating one factor matrix with another one fixed.

At the t-th iteration round, with W^t fixed, we get the first-order derivative of $F\left(W^t,H\right)$ with respect to H_{kj} as

$$\begin{aligned}
\frac{\partial F\left(W^t,H\right)}{\partial H_{kj}}
&= \sum_{i=1}^{m}\frac{\partial}{\partial H_{kj}}\sum_{q=1}^{n}\rho\left(V_{iq}-\left(W^tH\right)_{iq}\right)\\
&\quad + \frac{1}{2}\frac{\partial}{\partial H_{kj}}\left\|\Lambda H^l - Y^l\right\|_F^2\\
&= \sum_{i=1}^{m}\frac{\partial}{\partial H_{kj}}\left(\rho\left(V_{ij}-\left(W^tH\right)_{ij}\right)\right.\\
&\quad \left. + \sum_{q\neq j}\rho\left(V_{iq}-\left(W^tH\right)_{iq}\right)\right)\\
&\quad + \frac{1}{2}\frac{\partial}{\partial H_{kj}}\left(\mathrm{tr}\left(H^l\Lambda^T\Lambda H^l\right)-\mathrm{tr}\left(Y^{l^T}\Lambda H^l\right)\right.\\
&\quad \left. - \mathrm{tr}\left(H^{l^T}\Lambda^T Y^l\right)-\mathrm{tr}\left(Y^{l^T}Y^l\right)\right)_{kj}\\
&= \sum_{i=1}^{m}\frac{\left(V-W^tH\right)_{ij}}{\sqrt{1+\left(V-W^tH\right)^2_{ij}}}\left(-W^t_{ik}\right)\\
&\quad + \Lambda^T\Lambda H^l_{kj} - \Lambda^T Y^l_{kj}\\
&= \left(W^{t^T}\frac{W^tH}{\sqrt{1+\left(V-W^tH\right)^2}}\right)_{kj}\\
&\quad - \left(W^{t^T}\frac{V}{\sqrt{1+\left(V-W^tH\right)^2}}\right)_{kj}\\
&\quad + \Lambda^T\Lambda H^l_{kj} - \Lambda^T Y^l_{kj}.
\end{aligned} \tag{14}$$

Based on (14), with the gradient descent, the update rule for H_{kj} can be written as

$$\begin{aligned}
H^{t+1}_{kj} = H^t_{kj} - \alpha^t_{kj}&\left(\left(W^{t^T}\frac{W^tH^t}{\sqrt{1+\left(V-W^tH^t\right)^2}}\right)_{kj}\right.\\
&- \left(W^{t^T}\frac{V}{\sqrt{1+\left(V-W^tH^t\right)^2}}\right)_{kj}\\
&\left. + \Lambda^T\Lambda H^l_{kj} - \Lambda^T Y^l_{kj}\right),
\end{aligned} \tag{15}$$

where $\alpha^t_{kj} > 0$ is the step size. To preserve the non-negativity of H^{t+1}_{kj}, we adaptively set the step size as

$$\alpha^t_{kj} = \frac{H^t_{kj}}{\left(W^{t^T}\frac{W^tH^t}{\sqrt{1+\left(V-W^tH^t\right)^2}} + \Lambda^T\Lambda H^l - \Lambda^T Y^l\right)_{kj}}. \tag{16}$$

By substituting (16) into (15), we obtain the multiplicative update (MU) algorithm for H_{kj} as

$$H^{t+1}_{kj} = \frac{H^t_{kj}\left(W^{t^T}\frac{V}{\sqrt{1+\left(V-W^tH^t\right)^2}}\right)_{kj}}{\left(W^{t^T}\frac{W^tH^t}{\sqrt{1+\left(V-W^tH^t\right)^2}} + \Lambda^T\Lambda H^l - \Lambda^T Y^l\right)_{kj}}. \tag{17}$$

Algorithm 1 MUR Optimation for RSS-NMF

```
Input: $V \in \Re^{m\times n}, Y \in \Re^{r\times l}, 1 \le k \le r$
Output: $W \in \Re^{m\times r}, H \in \Re^{r\times n}$
Initialize $W_0, H_0, t = 0$.
Repeat:
    Update $H^{t+1}$ as
        $H^{t+1} = H^t \circ \frac{W^{t^T} \frac{V}{\sqrt{1+(V-W^tH^t)^2}}}{W^{t^T} \frac{W^tH^t}{\sqrt{1+(V-W^tH^t)^2}} + \Lambda^T\Lambda H^l - \Lambda^T Y^l}$.
    Update $W^{t+1}$ as
        $W^{t+1} = W^t \circ \frac{\left(\frac{V}{\sqrt{1+\left(V-W^tH^{t+1}\right)^2}}\right)H^{t+1^T}}{\left(\frac{W^tH^{t+1}}{\sqrt{1+\left(V-W^tH^{t+1}\right)^2}}\right)H^{t+1^T}}$.
    Update $\Lambda$ as
        $\Lambda_{kk} = \frac{\langle H^l_{k\cdot}, Y^l_{k\cdot}\rangle}{\langle H^l_{k\cdot}, H^l_{k\cdot}\rangle}, \forall 1 \le k \le r$.
    $t = t + 1$.
until Stopping criteria is met
```

The MU (17) can be written in the matrix form to update the whole factor matrix H as

$$H^{t+1} = H^t \circ \frac{W^{t^T} \frac{V}{\sqrt{1+(V-W^tH^t)^2}}}{W^{t^T} \frac{W^tH^t}{\sqrt{1+(V-W^tH^t)^2}} + \Lambda^T\Lambda H^l - \Lambda^T Y^l}. \tag{18}$$

We can derive the following MUR for W as

$$W^{t+1} = W^t \circ \frac{\left(\frac{V}{\sqrt{1+\left(V-W^tH^{t+1}\right)^2}}\right)H^{t+1^T}}{\left(\frac{W^tH^{t+1}}{\sqrt{1+\left(V-W^tH^{t+1}\right)^2}}\right)H^{t+1^T}}. \tag{19}$$

The update rules (18) and (19) are simple and easy to implement. We iteratively update W and H until the objective value of (12) does not change. The procedure is summarized in Algorithm 1.

Theorem 2 shows that (18) monotonically decreases the objective function of (12). We can similarly prove that (19) monotonically decreases the objective function of (12).

Theorem 2: Fixing W^t, when (15) updates H from H^t to H^{t+1}, the objective function of (12) monotonically decreases.

Similarly, we can prove that (17) decreases the objective function of (12). In summary, under the MUs (18) and (19), we have $F\left(W^{t+1}, H^{t+1}\right) \le F\left(W^t, H^{t+1}\right) \le F\left(W^t, H^t\right)$ for any $t \ge 0$.

C. FAST GRADIENT DESCENT FOR RSS-NMF

Although MUR guarantees decreasing the objective function of RSS-NMF, it does not guarantee convergence to any stationary point. The stationarity is important because it is necessary for finding a local minimum. Therefore, we proposed a fast gradient descent (FGD) algorithm for optimizing RSS-NMF. At the t-th iteration, the FGD algorithm updates both factor matrices as

$$\widehat{H}^{t+1} = H^t - \beta^{t+1} \frac{\overline{H}^t}{W^{t^T}\left(\frac{W^t\overline{H}^t}{\sqrt{1+\left(V-W^t\overline{H}^t\right)^2}}\right)+\delta} \circ \nabla_H F\left(W^t, H^t\right) \tag{20}$$

and

$$\widehat{W}^{t+1} = W^t - \gamma^{t+1} \frac{\overline{W}^t}{\left(\frac{\overline{W}^t\widehat{H}^{t+1}}{\sqrt{1+\left(V-\overline{W}^t\widehat{H}^{t+1}\right)^2}}\right)\widehat{H}^{t+1^T}+\delta} \circ \nabla_W F\left(W^t, \widehat{H}^{t+1}\right), \tag{21}$$

where δ is a positive constant, β^{t+1} and γ^{t+1} are step sizes for updating H and W, respectively, and both $\overline{H}^t$ and $\overline{W}^t$ are defined as

$$\overline{H}^t_{kj} = \begin{cases} H^t_{kj}, & \text{if } \left(\nabla_H F\left(W^t, H^t\right)\right)_{kj} \ge 0 \\ \max\left\{H^t_{kj}, \sigma\right\}, & \text{if } \left(\nabla_H F\left(W^t, H^t\right)\right)_{kj} < 0 \end{cases} \tag{22}$$

and

$$\overline{W}^t_{ik} = \begin{cases} W^t_{ik}, & \text{if } \left(\nabla_W F\left(W^t, \widehat{H}^{t+1}\right)\right)_{ik} \ge 0 \\ \max\left\{W^t_{ik}, \sigma\right\}, & \text{if } \left(\nabla_W F\left(W^t, \widehat{H}^{t+1}\right)\right)_{ik} < 0, \end{cases}$$

where σ is also a positive constant, and $\nabla_H F\left(W^t, H^t\right)$ and $\nabla_W F\left(W^t, \widehat{H}^{t+1}\right)$ are the derivatives of $F(W, H)$ with respect to W and H, respectively. The intermediate factor matrices $\widehat{H}^{t+1}$ and $\widehat{W}^{t+1}$ are normalized to obtain H^{t+1} and W^{t+1} as follows:

$$H^{t+1}_{k\cdot} = \begin{cases} \left\|\widehat{W}^{t+1}_{\cdot k}\right\|_1 \times \widehat{H}^{t+1}_{k\cdot}, & \text{if } \left\|\widehat{W}^{t+1}_{\cdot k}\right\|_1 > 0 \\ \widehat{H}^{t+1}_{k\cdot}, & \text{if } \left\|\widehat{W}^{t+1}_{\cdot k}\right\|_1 = 0 \end{cases} \tag{23}$$

and

$$W^{t+1}_{\cdot k} = \begin{cases} \widehat{W}^{t+1}_{\cdot k} / \left\|\widehat{W}^{t+1}_{\cdot k}\right\|_1, & \text{if } \left\|\widehat{W}^{t+1}_{\cdot k}\right\|_1 > 0 \\ \widehat{W}^{t+1}_{\cdot k}, & \text{if } \left\|\widehat{W}^{t+1}_{\cdot k}\right\|_1 = 0. \end{cases} \tag{24}$$

The modification does not introduce extra computational overheads. Therefore, the computational complexity of FGD is the same as that of MUR.

Looking more carefully at the FGD algorithm (20) - (21), we can find that they are intrinsically first-order gradient descent algorithms which search along the rescaled negative gradient direction. Here we proposed to search the optimal step size by the Newton algorithm.

At the t-th iteration round, fixing W^t, according to (20), the rescaled gradient of $F(W, H)$ with respect to H can be written as

$$\widehat{\nabla}_H F\left(W^t, H^t\right) = \frac{\overline{H}^t}{W^{t^T}\left(\frac{W^t\overline{H}^t}{\sqrt{1+\left(V-W^t\overline{H}^t\right)^2}}\right)+\delta} \circ \nabla_H F\left(W^t, H^t\right), \tag{25}$$

where $\nabla_H F\left(W^t, H^t\right) = W^{t^T}\left(\frac{W^tH^t-V}{\sqrt{1+(V-W^tH^t)^2}}\right)+\Lambda^T\Lambda H^l-\Lambda^T Y^l$ denotes the first-order derivative of $F\left(W^t, H\right)$ with respect to H. The update algorithm (20) can be written in the following gradient descent form

$$\widehat{H}^{t+1} = H^t - \beta^{t+1}\widehat{\nabla}_H F\left(W^t, H^t\right), \tag{26}$$

and the FGD procedure for optimizing RSS-NMF is summarized in Algorithm 2.

Algorithm 2 FGD Optimation for RSS-NMF

```
Input: $H^t \in \Re^{r\times n}, \beta^t$
Output: $H^{t+1} \in \Re^{r\times n}, \beta^{t+1}$
Initialize $\beta_0 = \beta^t, a = 0$.
Calculate $\overline{H}^t$ as (22).
Calculate $\widehat{\nabla}_H F\left(W^t, H^t\right)$ as (25).
Calculate $\lambda = \min\left\{\frac{W^t_{ik}}{(\widehat{\nabla}_W F(W^t,\widehat{H}^{t+1}))_{ik}} \middle| \left(\widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right)\right)_{ik} > 0\right\}$.
Set $\widehat{\beta}^t = 0.01 + 0.9 \times \lambda$.
Repeat:
    Update $\beta_{a+1}$ as
        $\beta_{a+1} = \beta_a - \frac{\phi_t'(\beta_a)}{\phi_t''(\beta_a)}$.
    $a = a + 1$.
until Stopping criteria is met
Set $\beta^{t+1} = \min\left\{\beta_a, \widehat{\beta}^t\right\}$.
Update $\widehat{H}^{t+1}$ as $\widehat{H}^{t+1} = H^t - \beta^{t+1}\widehat{\nabla}_H F\left(W^t, H^t\right)$.
Update $H^{t+1}$ according to (23).
```

Next, we will introduce how to determine a suitable step size β^t. To sufficiently decrease the objective function along the rescaled gradient direction, we solve the following line search problem

$$\min_{\beta\in D^t_H} F\left(W^t, H^t - \beta\widehat{\nabla}_H F\left(W^t, H^t\right)\right), \tag{27}$$

where the domain of β is set to

$$D^t_H = \left\{\beta | H^t - \beta\widehat{\nabla}_H F\left(W^t, H^t\right) \geq 0, \beta \geq 0\right\}$$

for preserving the non-negativity of $\widehat{H}^{t+1}$.

Let $\phi_t(\beta) = F\left(W^t, H^t - \beta\widehat{\nabla}_H F\left(W^t, H^t\right)\right)$, according to (20), we can obtain its first-order and second-order derivatives with respect to β as follows:

$$\begin{aligned}\phi_t'(\beta) &= \sum_{ij}\left(\frac{\left(W^t\widehat{\nabla}_H F\left(W^t, H^t\right)\right)_{ij}}{\sqrt{1 + \left(V - W^tH^t + \beta W^t\widehat{\nabla}_H F\left(W^t, H^t\right)\right)^2_{ij}}}\right.\\ &\quad\left.\times\frac{\left(V - W^tH^t + \beta W^t\widehat{\nabla}_H F\left(W^t, H^t\right)\right)_{ij}}{\sqrt{1 + \left(V - W^tH^t + \beta W^t\widehat{\nabla}_H F\left(W^t, H^t\right)\right)^2_{ij}}}\right)\end{aligned}$$

and

$$\phi_t''(\beta)=\sum_{ij}\frac{\left(W^t\widehat{\nabla}_H F\left(W^t, H^t\right)\right)^2_{ij}}{\left(1+\left(V-W^tH^t + \beta W^t\widehat{\nabla}_H F\left(W^t, H^t\right)\right)^2_{ij}\right)^{3/2}}.$$

Since $\phi_t(\beta)$ is continuous and differentiable, and its second-order derivative is non-negative, i.e., $\phi_t''(\beta) \geq 0$, the line search problem (27) is convex. It implies that there exists a global minimum of $\phi_t(\beta)$. We utilized the Newton algorithm to solve (27) as follows:

$$\beta_{a+1} = \beta_a - \frac{\phi_t'\left(\beta_a\right)}{\phi_t''\left(\beta_a\right)}, \tag{28}$$

where a is the iteration counter. The Newton algorithm (28) converges rapidly and obtains the minimal β_* of (27). Subsequently, we obtain the final step size as

$$\beta^{t+1} = \min\left\{\beta_*, \widehat{\beta}^t\right\}, \tag{29}$$

where $\widehat{\beta}^t = \tau \sup\left(D^t_H\right)$ and $\tau(0 < \tau < 1)$ is used to ensure that H^{t+1} is not too close to the boundary of the domain. The variable $\sup\left(D^t_H\right)$ can be computed as

$$\sup\left(D^t_H\right) = \min\left\{\frac{H^t_{kj}}{\left(\widehat{\nabla}_H F\left(W^t, H^t\right)\right)_{kj}} \middle| \left(\widehat{\nabla}_H F\left(W^t, H^t\right)\right)_{kj} > 0\right\}. \tag{30}$$

Since $1\in D^t_H$, we know that $\sup\left(D^t_H\right) \geq 1$.

Similarly, the optimal step size for updating W^{t+1} can also be searched by using the Newton algorithm. Fixing H^{t+1}, according to (21), the rescaled gradient direction of $F(W, H)$ with respect to W can be written as

$$\begin{aligned}\widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right) &= \frac{\overline{W}^t}{\left(\frac{\overline{W}^t\widehat{H}^{t+1}}{\sqrt{1+\left(V-\overline{W}^t\widehat{H}^{t+1}\right)^2}}\right)\widehat{H}^{t+1^T} + \delta}\\ &\quad\circ \nabla_W F\left(W^t, \widehat{H}^{t+1}\right),\end{aligned} \tag{31}$$

where $\nabla_W F\left(W^t, \widehat{H}^{t+1}\right) = \left(\frac{W^t\widehat{H}^{t+1}-V}{\sqrt{1+\left(V-W^t\hat{H}^{t+1}\right)^2}}\right)\widehat{H}^{t+1^T}$ denotes the first-order derivative of $F\left(W, \hat{H}^{t+1}\right)$ with respect to W. The update algorithm (21) can be written in the following gradient descent form

$$\widehat{W}^{t+1} = W^t - \gamma^{t+1}\widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right). \tag{32}$$

Next, we will introduce how to determine a suitable step size γ^{t+1}. To sufficiently decrease the objective function along the rescaled gradient direction, we solve the following line search problem

$$\min_{\beta\in D^t_W} F\left(W^t - \gamma\widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right), \widehat{H}^{t+1}\right), \tag{33}$$

where the domain of γ is set to

$$D^t_W = \left\{\gamma | W^t - \gamma\widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right) \geq 0, \gamma \geq 0\right\} \tag{34}$$

for preserving the non-negativity of $\widehat{W}^{t+1}$.

Let $\psi_t(\gamma) = F\left(W^t - \gamma \widehat{\nabla}_H F\left(W^t, \widehat{H}^{t+1}\right), H^t\right)$, according to (20), we can obtain its first-order and second-order derivatives with respect to as follows:

$$\psi_t'(\gamma) = \sum_{ij} \left(\frac{\left(\widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right) \widehat{H}^{t+1}\right)_{ij}}{\sqrt{1 + \begin{pmatrix} V - W^t \widehat{H}^{t+1} \\ +\gamma \widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right) \widehat{H}^{t+1} \end{pmatrix}_{ij}^2}} \times \frac{\left(V - W^t \widehat{H}^{t+1} + \gamma \widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right) \widehat{H}^{t+1}\right)_{ij}}{\sqrt{1 + \begin{pmatrix} V - W^t \widehat{H}^{t+1} \\ +\gamma \widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right) \widehat{H}^{t+1} \end{pmatrix}_{ij}^2}} \right)$$

and

$$\psi_t''(\gamma) = \sum_{ij} \frac{\left(\widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right) \widehat{H}^{t+1}\right)_{ij}^2}{\left(1 + \begin{pmatrix} V - W^t \widehat{H}^{t+1} \\ +\gamma \widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right) \widehat{H}^{t+1} \end{pmatrix}_{ij}^2\right)^{3/2}}.$$

Since $\psi_t(\gamma)$ is continuous and differentiable, and its second-order derivative is non-negative, i.e., $\psi_t''(\gamma) \geq 0$, the line search problem (33) is convex. It implies that there exists a global minimum of $\psi_t(\gamma)$. We utilized the Newton algorithm to solve (33) as follows:

$$\gamma_{b+1} = \gamma_b - \frac{\psi_t'\left(\gamma_b\right)}{\psi_t''\left(\gamma_b\right)}, \tag{35}$$

where b is the iteration counter. The Newton algorithm (35) converges rapidly and obtains the minimal γ_* of (33). Subsequently, we obtain the final step size as

$$\gamma^{t+1} = \min\left\{\gamma_*, \hat{\gamma}^t\right\}, \tag{36}$$

where $\hat{\gamma}^t = \tau \sup\left(D_W^t\right)$ and $\tau (0 < \tau < 1)$ is used to ensure that W^{t+1} is not too close to the boundary of the domain. The variable $\sup\left(D_W^t\right)$ can be computed as

$$\sup\left(D_W^t\right) = \min\left\{ \frac{W_{ik}^t}{\left(\widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right)\right)_{ik}} \,\middle|\, \left(\widehat{\nabla}_W F\left(W^t, \widehat{H}^{t+1}\right)\right)_{ik} > 0 \right\}. \tag{37}$$

Since $1 \in D_W^t$, we know that $\sup\left(D_W^t\right) \geq 1$.

The following section will prove that the FGD algorithm (20)-(24) with the Newton based line search (29) and (36) converges to a stationary point. If the point (H^*, W^*) is a stationary point, it is necessary to satisfy the following K.K.T. conditions:

$$\begin{aligned} H^* &\geq 0, & W^* &\geq 0, \\ \nabla_H\left(W^*, H^*\right) &\geq 0, & \nabla_W\left(W^*, H^*\right) &\geq 0, \\ H^* \circ \nabla_H\left(W^*, H^*\right) &= 0, & W^* \circ \nabla_W\left(W^*, H^*\right) &= 0. \end{aligned} \tag{38}$$

D. CONVERGENCE PROOF

The following Theorem 3 will prove that the FGD algorithm (20) - (24) with line search (29) and (36) converges to a stationary point. Before proving Theorem 3, we prove the following Lemma 1 to Lemma 4. For the clarity of presentation, we deduce the proofs of Lemma 1 to Lemma 4 to appendices.

Lemma 1: If the FGD algorithm (20) - (24) with the line search (29) and (36) generate an infinite sequence $\left\{H^t, W^t\right\}$, then

$$\begin{aligned} F\left(W^{t+1}, H^{t+1}\right) &= F\left(\widehat{W}^{t+1}, \widehat{H}^{t+1}\right) \\ &\leq F\left(W^t, \widehat{H}^{t+1}\right) \leq F\left(W^t, H^t\right). \end{aligned} \tag{39}$$

Lemma 2: Assume $\left\{H^t\right\}$, $t \in \mathrm{T}$, is a convergent subsequence and

$$\lim_{t \in \mathrm{T}, \mathrm{t} \to \infty} H^t = H^*. \tag{40}$$

Then

$$\lim_{t \in \mathrm{T}, \mathrm{t} \to \infty} \widehat{H}^{t+1} = H^*. \tag{41}$$

Next, we will prove that at any limit point $\{H^*, W^*\}$, the matrix H^* satisfies the K.K.T. conditions.

Lemma 3: Assume $\left\{H^t, W^t\right\}$, $t \in T$, is a convergent subsequence, and

$$\lim_{t \in T, t \to \infty} \left(H^t, W^t\right) = \left(H^*, W^*\right). \tag{42}$$

We have that: (1) if $H_{kj}^* > 0$, then $\nabla_H F\left(W^*, H^*\right)_{kj} = 0$; (2) if $H_{kj}^* = 0$, then $\nabla_H F\left(W^*, H^*\right)_{kj} \geq 0$.

Furthermore, the above conclusions (1) and (2) imply that any limit point of the sequence $\left\{H^t, W^t\right\}$ is a staionary point.

Lemma 4: The sequence $\left\{H^t, W^t\right\}$ generated by the FGD algorithm (20) - (24) has at least one limit point.

We are now ready to prove that the FGD algorithm (20) - (24) with the line search (29) and (36) converges to a stationary point.

Theorem 3: The sequence $\left\{H^t, W^t\right\}$ generated by the FGD algorithm (20) - (24) has at least one stationary point.

Proof: By **Lemma 2**, there exist a convergent subsequence such that

$$\lim_{t \in T, t \to \infty} \left(W^t, \widehat{H}^{t+1}\right) = \left(W^*, H^*\right). \tag{43}$$

Then, we can use the same proof procedure like **Lemma 3** to show the stationarity condition on W^*. Together with **Lemma 3**, we know that any limit point of the sequence $\left\{H^t, W^t\right\}$ is a stationary point. According to the **Lemma 4**, we know that the sequence $\left\{H^t, W^t\right\}$ has at least one limit point. Above all, the generated sequence $\left\{H^t, W^t\right\}$ has at least one stationary point. This completes the proof.

E. COMPUTATIONAL COMPLEXITY ANALYSIS

Since the object function of RSS-NMF and other typical NMF-based methods are minimized by alternating optimization algorithms, it is necessary to analyse their computational cost of one iteration round. We firstly discuss the computational complexity of RSS-NMF optimized by MUR and FGD in detail and then give the complexity analysis of other NMF-based methods in the experiments.

For MUR, inspired by [34], we compute $W^t H^{t+1} H^{t+1^T}$ as $W^t \left(H^{t+1} H^{t+1^T} \right)$, thus, reduce the computational complexity of (18) and (19) from $\mathcal{O}\left(mnr\right)$ to $\mathcal{O}\left(max\{m,n\}r^2\right)$ for $r \ll min\{m,n\}$. Thus, the overall complexity is $\mathcal{O}\left(mnr + max\{m,n\}r^2 + rn^2\right)$.

For FGD, the time cost is mainly spent on (20), (21) and (28). The complexity of (20) and (21) is same as (18) and (19) and the time cost of (28) is $\mathcal{O}\left(mnr + max\{m,n\}r^2 + rn^2\right)$. Therefore, the overall cost is the same as MUR for minimizing RSS-NMF in one iteration, i.e., $\mathcal{O}\left(mnr + max\{m,n\}r^2 + rn^2\right)$. Since FGD has a faster convergence speed compared with MUR, the overall time cost of FGD is less than MUR.

For one step, the overall cost for NMF is $\mathcal{O}\left(mnr\right)$ and the overall cost for CNMF in F-norm formulation is the same; the time cost for CNMF in divergence formulation is $\mathcal{O}\left(n(m+n)r\right)$; the complexity for GNMF is $\mathcal{O}\left(mnr + n^2 m\right)$; and for semi-GNMF, extra $\mathcal{O}\left(n^3\right)$ is needed to learn the Mahalanobis distance space [35], thus the overall cost is $\mathcal{O}\left(mnr + n^2 m + n^3\right)$. For single step, the time cost of RSS-NMF is a little more than NMF and CNMF in F-norm formulation. However, the overall comutational complexity of RSS-NMF may not be slower, since its iteration number for convergence is smaller as shown in the experiments below.

IV. EXPERIMENTS

In this section, we evaluate the performance of our proposed RSS-NMF comparing with nine other representative algorithms, including both unsupervised models and semi-supervised models. The unsupervised clustering methods include PCA [15], NMF [16], GNMF [17], IDEC [32] and RNMF; the semi-supervised clustering methods include Semi-GNMF, CNMF [18], MSAEClust [30], SSC-SR [33] and RSS-NMF.

Several comparison experiments are carried out for effectiveness evaluation of RSS-NMF on four image datasets including Yale, COIL-100, UMIST, and GT. In addition, We analyze the convergence speed of RSS-NMF with FGD versus MUR.

A. EVALUATION METRICS

Accuracy (AC) and normalized mutual information (NMI) are two important metrics widely used to evaluate the clustering performance of different clustering algorithms.

Accuracy (AC) is used to evaluate cluster results by comparing the obtained label of each sample with the label provided by the dataset. Given a dataset containing images, let l^* is the label obtained by applying different algorithms, and l is the ground true label. The accuracy (AC) is defined as

$$\mathrm{AC} = \frac{\sum_{i-1}^{n} \delta\left(l, \mathrm{map}\left(l^*\right)\right)}{n}, \tag{44}$$

TABLE 1. Statistics of the four datasets.

Dataset	Size (N)	Dimensionality (M)	No. of classes (K)
$Yale$	165	1024	15
$COIL-100$	7200	1024	100
$UMIST$	575	1600	20
GT	750	768	50

where $\delta(\mathrm{x}, \mathrm{y})$ is the delta function that equals one if $\mathrm{x} = \mathrm{y}$ and equals zero otherwise, and $\mathrm{map}\left(l^*\right)$ is the mapping function that maps cluster to the corresponding predicted label. The best mapping is gained by Kuhn-Munkres algorithm.

The other metric is normalized mutual information (NMI). Let C is the set of clusters gained from the dataset and C' is obtained by applying proposed algorithm. The mutual information $\mathrm{MI}\left(\mathrm{C}, \mathrm{C}'\right)$ is defined as

$$\mathrm{MI}\left(\mathrm{C}, \mathrm{C}'\right) = \sum_{c_i \in C, c_j \in C'} p\left(c_i, c'_j\right) \log \frac{p\left(c_i, c'_j\right)}{p\left(c_i\right) p\left(c'_j\right)}, \tag{45}$$

where $\mathrm{p}\left(c_i\right)$ and $\mathrm{p}\left(c'_j\right)$ are the probabilities that a image belongs to the cluster c_i and c'_j, respectively, and $\mathrm{p}\left(c_i, c'_j\right)$ denotes the joint probability that this arbitrarily selected image belongs to the cluster c_i as well as c'_j at the same time. $\mathrm{MI}\left(\mathrm{C}, \mathrm{C}'\right)$ takes values between zero and $\max\left(\mathrm{H(C)}, \mathrm{H}\left(C'\right)\right)$, where $\mathrm{H(C)}$ and $\mathrm{H(C')}$ are the entropies of C and C', respectively. It reaches the maximum $\max\left(\mathrm{H(C)}, \mathrm{H}\left(C'\right)\right)$ when the two sets of image clusters are identical and it becomes zero when the two sets are completely independent. In our experiment, we use the normalized mutual information (NMI) which takes values between zero and one. $\mathrm{NMI}\left(C, C'\right)$ is defined as

$$\mathrm{NMI}\left(C, C'\right) = \frac{MI\left(C, C'\right)}{\max\left(H(C), H\left(C'\right)\right)}, \tag{46}$$

B. CLUSTERING PERFORMANCE EVALUATION

We evaluate the clustering performance on four image datasets. The important statistics of these datasets are reported in Table 1 and the details are described individually.

For each dataset, we conduct the evaluations with different number of clusters k. We randomly select k categories from each dataset. By applying different algorithms as listed above to different datasets, we obtain new data representation H, the dimensionality of which is the same as the number of clusters k.

The label information is important for semi-supervised clustering algorithms, and with the increase of labeled data, the cluster results are better and more accurate. For each cluster, we assign labels for 10% and less samples as the labeled data at random. In our experiments, we randomly pick up 2, 2, 5, and 10 labeled images on the Yale, COIL-100, UMIST, and GT dataset for each cluster.

TABLE 2. The AC and NMI of RSS-NMF, PCA, NMF, GNMF, IDEC, RNMF, Semi-GNMF, CNMF, MSAEClust and SSC-SR on the Yale dataset.

k			2	5	8	10	15
Unsupervised algorithms	AC	PCA	63.64	52.67	36.18	37.53	30.45
		NMF	69.69	54.45	46.00	40.22	36.89
		GNMF	71.50	51.09	**49.67**	39.58	30.66
		IDEC	70.42	**61.53**	48.11	42.69	38.31
		RNMF	**77.67**	58.23	48.45	**44.08**	**40.71**
	NMI	PCA	45.38	33.68	41.66	39.02	44.10
		NMF	34.82	38.18	42.65	43.76	47.31
		GNMF	38.54	35.22	38.07	40.40	39.34
		IDEC	**46.26**	39.51	45.03	44.82	39.06
		RNMF	42.63	**44.56**	**48.91**	**47.59**	**50.63**
Semi-supervised algorithms	AC	Semi-GNMF	69.17	53.39	56.79	41.19	35.54
		CNMF	66.12	55.00	55.64	43.33	39.58
		MSAEClust	70.54	69.88	**63.21**	56.00	49.17
		SSC-SR	74.88	60.72	60.15	58.34	46.39
		RSS-NMF	**80.24**	**77.67**	60.11	**62.42**	**58.40**
	NMI	Semi-GNMF	39.03	37.79	39.88	41.77	42.10
		CNMF	33.67	42.56	45.20	45.50	48.27
		MSAEClust	43.63	39.50	42.62	38.04	39.03
		SSC-SR	**52.77**	44.03	43.97	44.82	38.55
		RSS-NMF	49.33	**44.26**	**46.42**	**49.57**	**55.92**

Then K-means is applied to the new data representation H for clustering. For a particular algorithm and dataset, clustering experiments is repeated ten times and the final result is obtained by averaging the test values. Finally, we compare the obtained clusters with the labels from dataset to compute the AC and NMI.

We run algorithms for n iterative rounds until convergence. The convergence criterion is defined as

$$\left|\frac{F_n - F_{n-1}}{F_n - F_1}\right| < \xi, \tag{47}$$

where ξ is a positive constant and empirically set to a small value, and F_n is the loss function value in the n-th iteration of each algorithm.

In our proposed RSS-NMF, there are four parameters, i.e., δ, σ, τ, and ξ. Parameters selection is a complex problem. Fortunately, performance of RSS-NMF is less sensitive to the above four parameters when they are in the range of $\left[10^{-6}, 10^{-9}\right]$, $\left[10^{-6}, 10^{-9}\right]$, $[0, 1]$, and $\left[10^{-4}, 10^{-7}\right]$, respectively. In our experiments, δ, σ, τ, and ξ is set to 10^{-6}, 10^{-6}, 0.9, and 10^{-6}, respectively. For other algorithms, the parameters are selected when they can achieve the highest performance.

C. PERFORMANCE EVALUATION AND COMPARISONS

Tables 2, 3, 4, and 5 show the detailed clustering accuracy and normalized mutual information of ten algorithms on Yale, COIL-100, UMIST, and GT dataset, respectively.

1) YALE DATASET

The Yale database [21] contains 165 grayscale images collected from 15 individuals. There are 11 images per subject with different facial expression or configuration. In our experiments, each image is normalized to 32×32 pixels with 256 gray levels per pixel.

Table 2 describes the accuracy and normalized mutual information of ten algorithms (including unsupervised versus semi-supervised algorithms) with different number of clusters k on the Yale dataset. It shows that our proposed RSS-NMF achieves better performance than other state-of-the-art algorithms significantly on the Yale dataset. Specifically, compared to the up-to-date semi-supervised clustering methods, e.g. MSAEClust, RSS-NMF raises the AC by 9.7, 7.79, 6.42, and 9.23 percent when k is set to 2, 5, 10, and 15.

2) COIL-100 DATASET

The COIL-100 database [25] contains 7200 images of 100 objects. The objects were placed on a motorized turntable against a black background. The turntable was rotated through 360 degrees to vary object pose with respect to a fixed color camera. Images of the objects were taken at pose intervals of 5 degrees. This corresponds to 72 poses per object. In our experiments, all images was normalized to 32×32 pixels.

Table 3 gives the clustering results on the COIL-100 dataset. By comparing the accuracy and normalized mutual information of RSS-NMF with other state-of-the-art models, it can be concluded that RSS-NMF outperforms other algorithms regardless of the number of clusters. Specifically, RSS-NMF gains 4 and 5 highest values in AC and NMI respectively. Furthermore, the NMF-based algorithms outperform the PCA algorithm, which indicates that the part-based representation enhances clustering effect compared with the global representation of data.

TABLE 3. The AC and NMI of RSS-NMF, PCA, NMF, GNMF, IDEC, RNMF, Semi-GNMF, CNMF, MSAEClust and SSC-SR on the COIL-100 dataset.

k			10	20	50	80	100
Unsupervised algorithms	AC	PCA	64.44	55.57	52.34	47.78	48.10
		NMF	64.36	58.44	49.98	50.99	41.12
		GNMF	84.11	63.41	54.86	50.21	45.65
		IDEC	**86.84**	69.06	51.37	52.86	48.33
		RNMF	85.22	**74.85**	**68.17**	**58.56**	**49.22**
	NMI	PCA	75.23	73.48	74.62	72.11	73.47
		NMF	77.50	79.31	63.24	74.08	69.40
		GNMF	84.73	74.37	76.50	70.55	**76.12**
		IDEC	86.20	**85.33**	79.05	**76.58**	71.44
		RNMF	**87.05**	78.86	**79.34**	72.46	74.68
Semi-supervised algorithms	AC	Semi-GNMF	92.08	66.78	60.41	47.46	41.35
		CNMF	80.21	70.41	68.80	50.19	44.07
		MSAEClust	**90.35**	78.48	73.08	60.77	49.63
		SSC-SR	85.79	79.05	73.49	64.70	50.13
		RSS-NMF	86.30	**79.16**	**77.68**	**65.81**	**53.14**
	NMI	Semi-GNMF	91.64	79.88	74.65	73.38	68.45
		CNMF	84.92	82.51	78.37	78.36	70.61
		MSAEClust	91.70	84.14	85.22	79.03	74.18
		SSC-SR	88.06	83.55	79.32	77.61	74.92
		RSS-NMF	**92.27**	**88.20**	**89.48**	**85.64**	**79.57**

TABLE 4. The AC and NMI of RSS-NMF, PCA, NMF, GNMF, IDEC, RNMF, Semi-GNMF, CNMF, MSAEClust and SSC-SR on the UMIST dataset.

k			2	5	8	10	20
Unsupervised algorithms	AC	PCA	69.13	68.81	62.87	55.46	41.08
		NMF	72.21	53.78	60.36	60.44	47.32
		GNMF	78.54	71.00	78.54	**69.54**	43.01
		IDEC	**80.42**	73.61	74.50	68.42	**64.46**
		RNMF	72.42	**76.45**	**81.33**	61.28	52.66
	NMI	PCA	79.46	61.62	70.39	68.33	55.34
		NMF	80.49	76.26	69.58	62.45	59.69
		GNMF	87.16	77.55	75.44	73.98	63.21
		IDEC	87.99	**84.28**	74.23	70.54	**69.66**
		RNMF	**89.34**	81.11	**77.21**	**79.60**	61.54
Semi-supervised algorithms	AC	Semi-GNMF	75.31	69.60	68.13	72.14	50.46
		CNMF	84.90	75.50	68.44	69.44	60.22
		MSAEClust	82.73	78.99	72.46	64.38	65.50
		SSC-SR	87.81	84.24	79.02	**73.18**	64.51
		RSS-NMF	**88.52**	**84.33**	**80.05**	71.40	**67.58**
	NMI	Semi-GNMF	77.10	70.41	71.43	54.52	68.33
		CNMF	91.33	80.54	79.65	75.74	60.45
		MSAEClust	88.73	84.42	85.90	78.22	68.00
		SSC-SR	**91.85**	86.44	86.30	**78.46**	69.58
		RSS-NMF	90.42	**87.07**	**88.41**	76.94	**69.81**

3) UMIST DATASET

The UMIST database [22] contains 575 of 20 distinct subjects taken in different poses from profile to frontal views. All images are down sampled to a size of 40 × 40 pixels and reshaped to a vector.

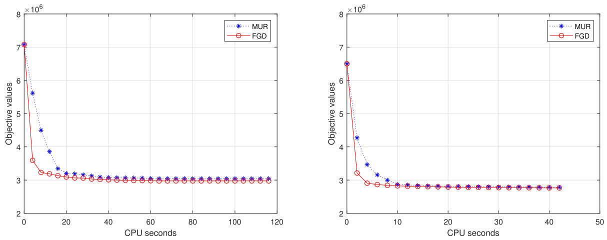
Table 4 shows that RSS-NMF and SSC-SR outperform the other algorithms. Although RNMF performs well for several clusters, its accuracy declines as the number of clusters becomes higher. According to experimental results, we notice that with the number of categories k varying from 2 to 20, both accuracy and normalized mutual information of all algorithms decrease. It is necessary to carry out sufficient experiments for each k each datasets, thus the clustering results are more compelling.

4) GT DATASET

The GT database [27] contains 750 images of 50 people taken in two or three sessions. All people in the database are represented by 15 color JPEG images with cluttered background taken at resolution 640 × 480 pixels. The pictures show frontal and/or tilted faces with different facial expressions, lighting conditions and scale. In our experiments, the average size of the faces images is normalized to 150×150 pixels.

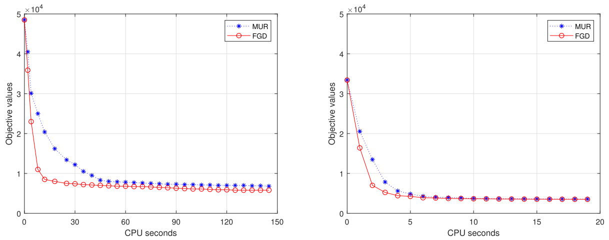
TABLE 5. The AC and NMI of RSS-NMF, PCA, NMF, GNMF, IDEC, RNMF, Semi-GNMF, CNMF, MSAEClust and SSC-SR on the GT dataset.

k			2	5	10	30	50
Unsupervised algorithms	AC	PCA	78.26	62.34	64.45	56.48	60.38
		NMF	71.95	63.90	66.38	52.20	54.77
		GNMF	76.08	**79.66**	58.07	63.44	63.36
		IDEC	78.92	77.53	**72.86**	62.77	60.38
		RNMF	**87.93**	75.40	67.72	**70.33**	**64.05**
	NMI	PCA	84.33	77.70	63.43	59.55	57.21
		NMF	79.95	77.84	69.97	61.03	60.48
		GNMF	84.07	74.41	67.00	60.36	71.36
		IDEC	82.73	83.64	77.49	76.08	**72.11**
		RNMF	**85.54**	**89.86**	**79.23**	**80.54**	65.71
Semi-supervised algorithms	AC	Semi-GNMF	71.33	63.24	66.36	65.44	56.98
		CNMF	70.66	68.35	60.01	70.49	54.33
		MSAEClust	79.56	72.82	68.04	69.22	63.66
		SSC-SR	82.16	**76.41**	69.07	68.29	66.18
		RSS-NMF	**89.33**	71.39	**70.62**	**73.40**	**69.38**
	NMI	Semi-GNMF	81.27	76.58	71.22	66.36	53.62
		CNMF	76.96	70.83	64.99	**79.68**	58.30
		MSAEClust	83.47	84.33	79.56	74.32	**70.66**
		SSC-SR	79.28	77.02	72.89	75.29	68.44
		RSS-NMF	**90.30**	**89.62**	**81.27**	77.95	64.86



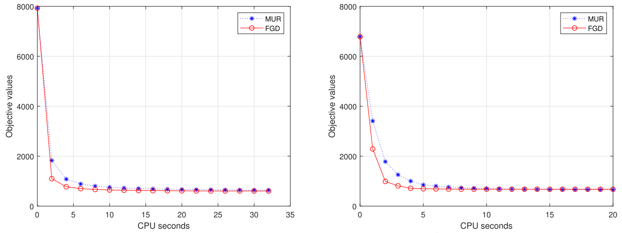
(a) r = 15 (b) r = 5

FIGURE 2. Objective values versus CPU seconds of MUR and FGD on the Yale dataset with reduced dimensionality. (a) r = 15. (b) r = 5.



(a) r = 100 (b) r = 10

FIGURE 3. Objective values versus CPU seconds of MUR and FGD on the COIL-100 dataset with reduced dimensionality. (a) r = 100. (b) r = 10.



(a) r = 20 (b) r = 10

FIGURE 4. Objective values versus CPU seconds of MUR and FGD on the UMIST dataset with reduced dimensionality. (a) r = 20. (b) r = 10.

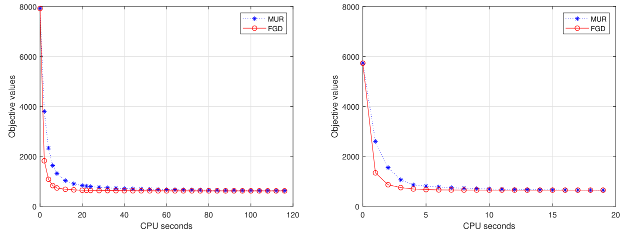
Table 5 presents the accuracy and normalized mutual information of ten algorithms on the UMIST dataset. Among unsupervised clustering algorithms, RNMF achieves the best performance. In semi-supervised algorithms, RSS-NMF performs the best in most cases. In contrast to unsupervised clustering algorithms, semi-supervised clustering algorithms, including Semi-GNMF, CNMF, MSAEClust, SSC-SR, RSS-NMF have obvious advantages in clustering accuracy and normalized mutual information. It suggests that label information plays a key role in improving discriminability of representations and clustering effect. Among semi-supervised clustering methods, our proposed RSS-NMF achieves the best effectiveness due to its robust formulation as presented in section III.

D. CONVERGENCE STUDY OF RSS-NMF WITH FGD VERSUS MUR

To evaluate the convergence efficiency of our proposed FGD, we compare FGD with MUR for optimizing RSS-NMF on Yale, COIL-100, UMIST, and GT database.

Figure 2 shows objective values versus CPU seconds of FGD and MUR on the Yale dataset with reduced dimensionality of 15 (see Figure 2(a)) and 5 (see Figure 2(b)). It shows that FGD reduces the objective values more rapidly and converges to a lower objective values compared with MUR.

Similarly, Figures 2, 3, and 4 present objective values versus CPU seconds of FGD and MUR on the COIL-100, UMIST, GT datasets with different reduced dimensionalities. For each test, the initialization and parameters selection of FGD and MUR are identical. Through a series of



(a) r = 50 (b) r = 10

FIGURE 5. **Objective values versus CPU seconds of MUR and FGD on the GT dataset with reduced dimensionality. (a) r = 50. (b) r = 10.**

comparison experiments, we observe FGD converges in less time and obtain smaller loss function values than MUR. Thus, we conclude that FGD works well for optimizing RSS-NMF and usually converges to a local minimum.

V. CONCLUSION

In this paper, we propose a robust semi-supervised NMF (RSS-NMF). RSS-NMF is robust to outliers by using the L_2/L_1-norm and incorporates few labeled examples by utilizing the structured normalization. In this way, RSS-NMF can learn from labeled and unlabeled data and obtains a more discriminating power representation space for data. In order to optimize RSS-NMF, We firstly introduce an MUR algorithm and theoretically show its convergence, and then propose an FGD algorithm to accelerate MUR. We also prove the convergence of FGD, and show that they converge to a stationary point. Empirical studies show that FGD is more efficient than MUR in optimizing RSS-NMF. We experimentally verify that RSS-NMF outperforms both unsupervised and semi-supervised models in terms of clustering performance.

APPENDIX A
PROOF OF THEOREM 1

Proof: For the convenience of presentation, we construct an auxiliary function $f(x) = \sqrt{1+x^2} - 1 - |x|$. Since the absolute function is non-differentiable at zero, we prove in two cases: (i) if $x < 0$, we have $f'(x) = \frac{x}{\sqrt{1+x^2}} + 1 = 1 - \sqrt{\frac{x^2}{1+x^2}} > 0$. Thus, $f(x)$ is an increasing function in the range $(-\infty, 0)$, and $f(-\infty) < f(x) < f(0)$; and (ii) if $x > 0$, we have $f'(x) = \frac{x}{\sqrt{1+x^2}} - 1 = \sqrt{\frac{x^2}{1+x^2}} - 1 < 0$. Thus, $f(x)$ is a decreasing function in the range $(0, +\infty)$, and $f(+\infty) < f(x) < f(0)$.

Since $\lim_{x\to+\infty} \sqrt{1+x^2} - x = 0$, we know that $f(+\infty) = \lim_{x\to+\infty} \sqrt{1+x^2} - 1 - x = -1$ and $f(-\infty) = \lim_{x\to-\infty} \sqrt{1+x^2} - 1 + x = \lim_{x\to+\infty} \sqrt{1+x^2} - 1 - x = -1$. By summarizing the above two cases and the additional case when $x = 0$, we know that $-1 \leq f(x) \leq 0$. Therefore, we have $\sum_{i=1}^{m} \rho(x_i) \leq ||x||_1$ and $||x||_1 - \sum_{i=1}^{m} \rho(x_i) \leq m$. This completes the proof.

APPENDIX B
PROOF OF THEOREM 2

Proof: Let $M_{ij}^t = \frac{1}{\sqrt{1+(V-W^tH^t)_{ij}^2}}$, we have that

$$\begin{aligned}
&\sum_{ij} \rho\left(\left(V - W^tH^{t+1}\right)_{ij}\right) - \sum_{ij} \rho\left(\left(V - W^tH^t\right)_{ij}\right) \\
&\quad - \frac{1}{2}\left(\sum_{ij} M_{ij}^t \left(V - W^tH^{t+1}\right)_{ij}^2\right. \\
&\quad \left. - \sum_{ij} M_{ij}^t \left(V - W^tH^t\right)_{ij}^2\right) \\
&= \sum_{ij}\left(\sqrt{1+\left(V - W^tH^{t+1}\right)_{ij}^2} - \sqrt{1+(V-W^tH^t)_{ij}^2}\right. \\
&\quad \left. - \frac{\left(V - W^tH^{t+1}\right)_{ij}^2}{2\sqrt{1+(V-W^tH^t)_{ij}^2}} + \frac{\left(V - W^tH^t\right)_{ij}^2}{2\sqrt{1+(V-W^tH^t)_{ij}^2}}\right) \\
&= \sum_{ij} \frac{\left(\begin{array}{c}\sqrt{1+\left(V - W^tH^{t+1}\right)_{ij}^2} \\ -\sqrt{1+(V-W^tH^t)_{ij}^2}\end{array}\right)^2}{-2\sqrt{1+(V-W^tH^t)_{ij}^2}} \\
&\leq 0. \qquad (48)
\end{aligned}$$

Since the previous H^t is fixed, the matrix M^t is also fixed, the MUR algorithm (18) intrinsically optimizes the weighted non-negative least squares model [1], i.e.,

$$\min_{H\geq 0} \sum_{ij} M_{ij}^t (V - W^tH)_{ij}^2. \qquad (49)$$

According to [1], we know that the objective function decreases under the MUR (18), i.e.,

$$\sum_{ij} M_{ij}^t (V - W^tH^{t+1})_{ij}^2 \leq \sum_{ij} M_{ij}^t (V - W^tH^t)_{ij}^2. \qquad (50)$$

By combining (48) and (50), we have

$$\sum_{ij} \rho((V - W^tH^{t+1})_{ij}) \leq \sum_{ij} \rho((V - W^tH^t)_{ij}). \qquad (51)$$

This completes the proof.

APPENDIX C
PROOF OF LEMMA 1

Proof: The first equality is strict as the normalization (23) and (24) does not change the objective value. We concentrate on proving the next two inequalities. Taking the last inequality for example, we prove at the following two steps:

(1) When the step size $\beta_H^t = 1$, let the updated value of (20) to be

$$\widetilde{H}^{t+1} = H^t - \frac{\overline{H}^t}{W^{t^T}\left(\frac{W^t\overline{H}^t}{\sqrt{1+(V-W^t\overline{H}^t)^2}}\right) + \delta} \circ \nabla_H F(W^t, H^t). \qquad (52)$$

At this step, we prove that $F\left(W^t, \widetilde{H}^{t+1}\right) \leq F\left(W^t, H^t\right)$. Let $M_{ij}^t = \frac{1}{\sqrt{1+(V-W^tH^t)_{ij}^2}}$, we consider the following formulation, i.e.,

$$\begin{aligned} &F\left(W^t, \widetilde{H}^{t+1}\right) - F\left(W^t, H^t\right) \\ &\quad - \frac{1}{2}\left(\sum_{ij} M_{ij}^t \left(V - W^t\widetilde{H}^{t+1}\right)_{ij}^2 \right. \\ &\quad \left. - \sum_{ij} M_{ij}^t \left(V - W^tH^t\right)_{ij}^2\right) \\ &= \sum_{ij} \rho\left(\left(V - W^t\widetilde{H}^{t+1}\right)_{ij}\right) - \sum_{ij} \rho\left(\left(V - W^tH^t\right)_{ij}\right) \\ &= \sum_{ij} \sqrt{1 + \left(V - W^t\widetilde{H}^{t+1}\right)_{ij}^2} - \sum_{ij} \sqrt{1 + (V - W^tH^t)_{ij}^2} \\ &\quad - \frac{1}{2}\left(\sum_{ij} \frac{\left(V - W^t\widetilde{H}^{t+1}\right)_{ij}^2}{\sqrt{1 + (V - W^tH^t)_{ij}^2}} \right. \\ &\quad \left. - \sum_{ij} \frac{\left(V - W^tH^t\right)_{ij}^2}{\sqrt{1 + (V - W^tH^t)_{ij}^2}}\right) \\ &= \sum_{ij} \frac{\left(\sqrt{1 + \left(V - W^t\widetilde{H}^{t+1}\right)_{ij}^2} - \sqrt{1 + (V - W^tH^t)_{ij}^2}\right)^2}{-2\sqrt{1 + (V - W^tH^t)_{ij}^2}} \\ &\leq 0, \end{aligned} \tag{53}$$

where the last equality is derived in a similar way as (48). From (53), we have

$$\begin{aligned} F(W^t, \widetilde{H}^{t+1}) - F(W^t, H^t) \leq \frac{1}{2}&\left(\sum_{ij} M_{ij}^t (V - W^t\widetilde{H}^{t+1})_{ij}^2\right. \\ &\left. - \sum_{ij} M_{ij}^t (V - W^tH^t)_{ij}^2\right). \end{aligned} \tag{54}$$

Therefore, it suffices to prove the following inequality

$$\frac{1}{2}\left(\sum_{ij} M_{ij}^t (V - W^t\widetilde{H}^{t+1})_{ij}^2 - \sum_{ij} M_{ij}^t (V - W^tH^t)_{ij}^2\right) \leq 0. \tag{55}$$

Towards this end, we divide the inequality into n separate inequalities because it is the sum of n functions, each of which relates to one column of H. Hence, it suffices to consider any column h and prove the following inequality

$$\frac{1}{2}\left(\sum_i m_i^t (v - W^t\widetilde{h}^{t+1})_i^2 - \sum_i m_i^t (v - W^th^t)_i^2\right) \leq 0, \tag{56}$$

where m^t,v, $\widetilde{h}^t$ and h^t are the corresponding columns of M^t, V, $\widetilde{H}^{t+1}$, and H^t, respectively.

Let $f(h) = \frac{1}{2} \parallel m^t \circ (v - W^th) \parallel_2^2$, we obtain the partial derivative of $f(h)$ at h^t as $\nabla_h f(h^t) = {W^t}^T D_m^t (W^t h^t - v)$, where $D_m^t = diag(m^t)$ is a diagonal matrix that puts the elements of m^t in the diagonal entries. According to (22), we have

$$\overline{h}_k^t = \begin{cases} h_k^t, & \text{if } \nabla_h f\left(h^t\right)_k \geq 0 \\ \max\left\{h_k^t, \sigma\right\}, & if\ \nabla_h f\left(h^t\right)_k < 0. \end{cases} \tag{57}$$

Let

$$\begin{aligned} I &= \{k | h_k^t > 0, \nabla_h f(h^t)_k \neq 0 \ or\ h_k^t = 0, \nabla_h f(h^t)_k < 0\} \\ &= \{k | \overline{h}_k^t > 0, \nabla_h f(h^t)_k \neq 0\} \end{aligned}$$

denote the indices of the elements of h^t that do not satisfy the K.K.T. conditions (38).

Define an auxiliary function

$$\begin{aligned} G\left(h, h^t\right) =& f\left(h^t\right) + \left(h - h^t\right)_I^T \nabla_h f\left(h^t\right)_I \\ &+ \frac{1}{2}\left(h - h^t\right)_I^T D\left(h^t, m^t, I\right)_{II} \left(h - h^t\right)_I, \end{aligned} \tag{58}$$

where $D(h^t, m^t, I)$ is a diagonal matrix whose diagonal entries are defined as

$$D\left(h^t, m^t, I\right)_{kk} = \begin{cases} \frac{\left({W^t}^T D_m^t W^t \overline{h}^t\right)_k + \delta}{\overline{h}_k^t}, & \text{if } k \in I \\ 0, & \text{if } k \notin I. \end{cases} \tag{59}$$

Since $D(h^t, m^t, I)_{II}$ is positive definite, $G(h, h^t)$ is a strictly convex function of h_I, and the unique minimal of $G(h, h^t)$ satisfies

$$D(h^t, m^t, I)_{II}(h - h^t)_I + \nabla_h f(h^t)_I = 0. \tag{60}$$

From (60), we have

$$\begin{aligned} &\underset{h_I}{\text{argmin}} G\left(h, h^t\right) \\ &\quad = h_I^t - D\left(h^t, m^t, I\right)_{II}^{-1} \nabla_h f\left(h^t\right)_I \\ &\quad = h_I^t - \frac{\overline{h}_I^t}{\left({W^t}^T D_m^t W^t \overline{h}^t\right)_I + \delta} \circ \nabla_h f\left(h^t\right)_I \\ &\quad = \tilde{h}_I^{t+1}. \end{aligned} \tag{61}$$

Let $I' = \{1, \ldots, r\}/I = \{k | \overline{h}_k^t = 0 \ or\ \nabla_h f(h^t)_k = 0\}$ denote the indices of h excluding I.According to (22), it is obvious that $\widetilde{h}_{I'}^{t+1} = h_{I'}^t$.

Since f(h) can be written as a quadratic function

$$\begin{aligned} f(h) =& f\left(h^t\right) + \left(h - h^t\right)^T \nabla_h f\left(h^t\right) \\ &+ \frac{1}{2}\left(h - h^t\right)^T {W^t}^T D_m^t W^t \left(h - h^t\right). \end{aligned} \tag{62}$$

For any h with $h_{I'} = h_{I'}^t$, we consider the following function

$$\begin{aligned} G\left(h, h^t\right) - f(h) =& \left(h - h^t\right)_I^T \nabla_h f\left(h^t\right)_I \\ &+ \frac{1}{2}\left(h - h^t\right)_I^T D\left(h^t, m^t, I\right)_{II} \left(h - h^t\right)_I \end{aligned}$$

$$\begin{aligned}
&- \left(h-h^t\right)^T \nabla_h f\left(h^t\right) \\
&- \frac{1}{2}\left(h-h^t\right)^T W^{t^T} D_m^t W^t \left(h-h^t\right) \\
= &\ \frac{1}{2}\left(h-h^t\right)_I^T \left(D\left(h^t, m^t, I\right)\right. \\
&\left. - W^{t^T} D_m^t W^t\right))_{II} \left(h-h^t\right)_I . \qquad (63)
\end{aligned}$$

Since, for any $k \in I$, we have

$$\begin{aligned}
&\left(D\left(h^t, m, I\right) - W^{t^T} D_m^t W^t\right)_{kk} \\
&= \frac{\left(W^{t^T} D_m^t W^t \overline{h}^t\right)_k + \delta}{\overline{h}_k^t} - \left(W^{t^T} D_m^t W^t\right)_{kk} \\
&= \frac{\left(W^{t^T} D_m^t W^t \overline{h}^t\right)_k + \delta - \left(W^{t^T} D_m^t W^t\right)_{kk} \overline{h}_k^t}{\overline{h}_k^t} \\
&> 0, \qquad (64)
\end{aligned}$$

where the matrix $(D(h^t, m^t, I) - W^{t^T} D_m^t W^t)_{II}$ is positive definite. Thus, according to (63), we have

$$G\left(h, h^t\right) - f(h) \geq 0. \qquad (65)$$

By substituting $\widetilde{h}^{t+1}$ into (65), we have

$$f(\widetilde{h}^{t+1}) \leq G(\widetilde{h}^{t+1}, h^t) \leq G(h^t, h^t) = f(h^t), \qquad (66)$$

where the second inequality is derived by (61). It implies that (55) is satisfied.

Therefore, we have $\frac{1}{2}(\sum_{ij} M_{ij}^t (V - W^t \widetilde{H}^{t+1})_{ij}^2 - \sum_{ij} M_{ij}^t (V - W^t H^t)_{ij}^2) \leq 0$. By (54), we have that $F(W^t, \widetilde{H}^{t+1}) \leq F(W^t, H^t)$.

(2) When the step size is determined by the line search procedure (29), we prove

$$F(W^t, \widehat{H}^{t+1}) \leq F(W^t, \widetilde{H}^{t+1}). \qquad (67)$$

Using the definition in Section 4.1, (67) implies that $\phi_t(\beta^t) \leq \phi_t(1)$. According to (29), we prove (67) in two scenarios: (i) If $\beta^t = \beta_*$, since β_* is the minimizer of (27), we know that $\phi_t(\beta^t) \leq \phi_t(1)$; and (ii) If $\beta^t = \widehat{\beta^t}$, we know that $\widehat{\beta^t} \leq \beta_*$ immediately. Since $\widehat{\beta^t} \geq 1$, there exist $0 \leq \mu \leq 1$, such that $\widehat{\beta^t} = \mu\beta_* + (1-\mu)$. Since ϕ_t is convex, by the Jensen inequality, we have $\phi_t(\widehat{\beta^t}) \leq \mu\phi_t(\beta_*) + (1-\mu)\phi_t(1) \leq \phi_t(1)$. Therefore, $F(\widehat{W}^{t+1}, \widehat{H}^{t+1}) \leq F(W^t, \widetilde{H}^{t+1})$.

In summary, we have that $F\left(\widehat{W}^{t+1}, \widehat{H}^{t+1}\right) \leq F\left(W^t, \widehat{H}^{t+1}\right)$. Similarly, we can prove that $F(\widehat{W}^{t+1}, \widehat{H}^{t+1}) \leq F(W^t, \widehat{H}^{t+1})$. This completes the proof.

APPENDIX D
PROOF OF LEMMA 2

Proof Lemma 1: and the property $F(W, H) \geq 0$ imply that $\{F(W^t, H^t)\}$ is a bounded decreasing sequence. Therefore, $\{F(W^t, H^t)\}$ converges globally, i.e.,

$$\lim_{t \in \mathrm{T}, t \to \infty} \left| F\left(W^t, \widehat{H}^{t+1}\right) - F\left(W^t, H^t\right) \right| = 0. \qquad (68)$$

We prove with contradiction. Assume this lemma is wrong, there exist an entry $(k.j)$ of H^t, a value $\varepsilon > 0$, and an infinite subset $\widehat{T} \subset T$ such that, for any $t \in \widehat{T}$,

$$|\widehat{H}_{kj}^{t+1} - H_{kj}^*| \geq \varepsilon. \qquad (69)$$

By the hypothesis, there is an infinite subset $\widetilde{T} \subset \widehat{T}$ such that, for any $t \in \widetilde{T}$,

$$|H_{kj}^t - H_{kj}^*| \leq \frac{\varepsilon}{2}. \qquad (70)$$

Combing (69) and (70), we know that, for any $t \in \widetilde{T}$,

$$|\widehat{H}_{kj}^{t+1} - H_{kj}^*| = \beta^t |\widetilde{H}_{kj}^{t+1} - H_{kj}^t| \geq \frac{\varepsilon}{2}. \qquad (71)$$

Since $\overline{H}_{kj}^t = 0$ implies $\widehat{H}_{kj}^{t+1} = H_{kj}^t$ according to (52), which will violate (71), we know that $\overline{H}_{kj}^t \neq 0$.

According to (67) and (54), we have

$$\begin{aligned}
F\left(W^t, \widehat{H}^{t+1}\right) - F\left(W^t, H^t\right) \leq \frac{1}{2}\Bigg(&\sum_{ij} M_{ij}^t \left(V - W^t \widetilde{H}^{t+1}\right)_{ij}^2 \\
&- \sum_{ij} M_{ij}^t \left(V - W^t H^t\right)_{ij}^2 \Bigg). \qquad (72)
\end{aligned}$$

Since $F(W^t, \widetilde{H}^{t+1}) - F(W^t, H^t)$ is the sum of separate column and the difference at each column is bounded. Assume h^t to be any column of H^t and $\widetilde{h}^{t+1}$ is the corresponding column of $\widetilde{H}^{t+1}$, then we know that

$$\widetilde{h}^{t+1} = h^t - \frac{\overline{h}^t}{W^{t^T} D_m^t W^t \overline{h}^t + \delta} \circ \nabla_h f(h^t). \qquad (73)$$

Therefore, by the definition of $D(h^t, m^t, I)_{II}$ given in (59), we have $\nabla_h f(h^t)_I = -D(h^t, m^t, I)_{II} (\widetilde{h}^{t+1} - h^t)_I$.

$$\begin{aligned}
&f(\widetilde{h}^{t+1}) - f(h^t) \\
&\quad \leq G(\widetilde{h}^{t+1}, h^t) - G(h^t, h^t) \\
&\quad = (\widetilde{h}^{t+1} - h^t)_I^T \nabla_h f(h^t)_I \\
&\qquad + \frac{1}{2}(\widetilde{h}^{t+1} - h^t)_I^T D(h^t, m^t, I)_{II} (\widetilde{h}^{t+1} - h^t)_I \\
&\quad = -\frac{1}{2}(\widetilde{h}^{t+1} - h^t)_I^T D(h^t, m^t, I)_{II} (\widetilde{h}^{t+1} - h^t)_I . \qquad (74)
\end{aligned}$$

Taking the inequality (74) over all the columns of $\widetilde{H}^{t+1}$ and H^t, we have

$$\begin{aligned}
&\sum_{ij} M_{ij}^t \left(V - W^t \widetilde{H}^{t+1}\right)_{ij}^2 - \sum_{ij} M_{ij}^t \left(V - W^t H^t\right)_{ij}^2 \\
&\quad \leq -\frac{1}{2}\left(\widetilde{H}_{\cdot j}^{t+1} - H_{\cdot j}^t\right)_I^T D\left(H_{\cdot j}^t, M_{\cdot j}^t, I\right)_{II} \left(\widetilde{H}_{\cdot j}^{t+1} - H_{\cdot j}^t\right)_I \\
&\quad = -\frac{1}{2}\sum_{i \in I} D\left(H_{\cdot j}^t, M_{\cdot j}^t, I\right)_{ii} \left(\widetilde{H}_{ij}^{t+1} - H_{ij}^t\right)^2 \\
&\quad \leq -\frac{1}{2} D\left(H_{\cdot j}^t, M_{\cdot j}^t, I\right)_{kk} \left(\widetilde{H}_{kj}^{t+1} - H_{kj}^t\right)^2
\end{aligned}$$

$$\begin{aligned}
&\leq -\frac{1}{2}\frac{\left(\widetilde{H}^{t+1}_{kj} - H^t_{kj}\right)^2 \delta}{\overline{H}^t_{kj}} \\
&\leq -\frac{1}{2}\frac{\left(\widetilde{H}^{t+1}_{kj} - H^t_{kj}\right)^2 \delta}{\max\left(H^t_{kj}, \sigma\right)} \\
&\leq 0. \qquad (75)
\end{aligned}$$

Taking the limit of the inequality (75), by (68) and (72), we have

$$\lim_{t\in\bar{T},t\to\infty} \tilde{H}^{t+1}_{kj} - H^t_{kj} = 0. \qquad (76)$$

It is obvious that (76) is a contradiction to (71). So, (68) is correct. This completes the proof.

APPENDIX E
PROOF OF LEMMA 3

Proof: According to the definition (22), we know that

$$\bar{H}^t_{kj} = \max\left(H^t_{kj}, \sigma\right) \parallel \bar{H}^t_{kj} = H^t_{kj}. \qquad (77)$$

So the sequence $\left\{\bar{H}^t_{kj}\right\}_{t\in T}$ may have two limit points, i.e., H^*_{kj} and σ. Since the number of indices (k, j) is finite, there is an infinite set $\bar{T} \subset T$ such that the sequence $\{\bar{H}^t\}_{t\in T}$ has limit point, i.e.,

$$\bar{H}^* = \lim_{t\in\bar{T},t\to\infty} \bar{H}^t. \qquad (78)$$

From **Lemma 2** and (20), we have

$$\begin{aligned}
&\lim_{t\in\bar{T},t\to\infty} H^t_{kj} - \hat{H}^{t+1}_{kj} \\
&\quad = \beta^* \frac{\left(\bar{H}^*\right)_{kj} \circ \nabla_H F(W^*, H^*)_{kj}}{\left(W^{*T}\left(\frac{W^*\bar{H}^*}{\sqrt{1+\left(V-W^*\bar{H}^*\right)^2}}\right)\right)_{kj} + \delta} = 0. \qquad (79)
\end{aligned}$$

From the definition (22), we know that $\bar{H}^*_{kj} \geq H^*_{kj}$. If $H^*_{kj} > 0$, then $\bar{H}^*_{kj} > 0$, and thus $\nabla_H F(W^*, H^*)_{kj} = 0$ by (79) and $\beta^* \geq 1$. It proves the first conclusion.

We prove the second conclusion by contradiction. Assume (2) is wrong. There exist (k, j) such that

$$H^*_{kj} = 0 \&\& \nabla_H F\left(W^*, H^*\right)_{kj} < 0. \qquad (80)$$

By the definition (22) and $\nabla_H F(W^*, H^*)_{kj} < 0$, for sufficiently large t , we have

$$\lim_{t\in\bar{T},t\to\infty} \bar{H}^t_{kj} = \bar{H}^*_{kj} = \sigma. \qquad (81)$$

Therefore, by $\beta^* \geq 1$, we have

$$\begin{aligned}
&\lim_{t\in\bar{T},t\to\infty} H^t_{kj} - \hat{H}^{t+1}_{kj} \\
&\quad = \beta^* \frac{\bar{H}^*_{kj} \circ \nabla_H F(W^*, H^*)_{kj}}{\left(W^{*T}\left(\frac{W^*\bar{H}^*}{\sqrt{1+\left(V-W^*\bar{H}^*\right)^2}}\right)\right)_{kj} + \delta} < 0. \qquad (82)
\end{aligned}$$

The formulation (82) is contradict with **Lemma 2**. Therefore, the second conclusion is also correct. This completes the proof.

APPENDIX F
PROOF OF LEMMA 4

Proof: It suffices to prove that $\left\{H^t, W^t\right\}$ are in a compact set, i.e., bounded and closed set. According to the structured normalization on $H^{(l)}$, at the t-th iteration round, the coefficients of labeled samples are

$$H^{(l),t}_{kj} \leftarrow \frac{H^{(l),t}_{k\cdot}, L^{(l)}_{k\cdot}}{H^{(l),t}_{k\cdot}, H^{(l),t}_{k\cdot}} H^{(l),t}_{kj} = \frac{\sum_{i\in C_k} H^{(l),t}_{ki} H^{(l),t}_{kj}}{\sum_{i=1}^{l} \left(H^{(l),t}_{ki}\right)^2}, \qquad (83)$$

where $C_k = \left\{i | L^{(l)}_{ki} = 1\right\}$. Since $H^{(l),t}_{ki} H^{(l),t}_{kj} \leq \frac{1}{2}\left(\left(H^{(l),t}_{ki}\right)^2 + \left(H^{(l),t}_{kj}\right)^2\right)$, Eq. (83) implies that

$$H^{(l),t}_{kj} \leftarrow \frac{\sum_{i\in C_k} H^{(l),t}_{ki} H^{(l),t}_{kj}}{\sum_{i=1}^{l} \left(H^{(l),t}_{ki}\right)^2} \leq \frac{|C_k|}{2}. \qquad (84)$$

Eq. (84) implies that $\left\{H^{(l),t}\right\}$ is bounded. The remainder thing is to show that $\left\{W^t\right\}$ and $\left\{H^{(u),t}\right\}$ are also bounded.

If $\left\{W^t\right\}$ is unbounded, there is a component W_{ik} and an infinite index set $\mathcal{T}$ such that

$$\lim_{t\in\mathcal{T},\, t\to\infty} W^t_{ik} \to \infty, \quad W^t_{ik} < W^{t+1}_{ik}, \ \forall t \in \mathcal{T}, \qquad (85)$$

and

$$\lim_{t\in\mathcal{T},\, t\to\infty} H^{(l),t}_{kj} = H^{(l),*}_{kj} \mathit{exists}, \quad \forall 1 \leq j \leq l. \qquad (86)$$

There must be $H^{(l),*}_{kj} = 0$ for any $1 \leq j \leq l$. Otherwise, there is an index j such that

$$\lim_{t\in\mathcal{T},\, t\to\infty} \left(W^t H^{(l),t}\right)_{ij} \geq \lim_{t\in\mathcal{T},\, t\to\infty} W^t_{ik} H^{(l),t}_{kj} = \infty. \qquad (87)$$

Then

$$\begin{aligned}
\lim_{t\to\infty} F\left(W^t, H^t\right) &\geq \lim_{t\to\infty} F\left(W^t, H^{(l),t}\right) \\
&\geq \lim_{t\in\mathcal{T},\, t\to\infty} \sum_{ij} \rho\left(V_{ij} - \left(W^t H^{(l),t}\right)_{ij}\right) \\
&= \infty. \qquad (88)
\end{aligned}$$

Since (88) is contradicted with **Lemma 1**, which shows that $F\left(W^t, H^t\right)$ is decreasing. Since $H^{(l),*}_{kj} = 0, \forall 1 \leq j \leq l$, we have

$$H^{(l),t}_{kj} = 0, \quad \forall 1 \leq j \leq l, \ \forall t \in \mathcal{T} \ \mathit{large\ enough}. \qquad (89)$$

Then

$$\nabla_W F\left(W^t, H^{(l),t}\right)_{ik} = 0, \quad \forall t \in \mathcal{T}. \qquad (90)$$

So,

$$\hat{W}^{t+1}_{ik} = W^t_{ik}, \quad \forall t \in \mathcal{T}. \qquad (91)$$

By (89), we have

$$\lim_{t\in\mathcal{T},\, t\to\infty} \hat{H}_{kj}^{(l),t+1} = \lim_{t\in\mathcal{T},\, t\to\infty} H_{kj}^{(l),t} = 0, \quad \forall 1 \le j \le l. \tag{92}$$

Therefore, in structured normalization, $\hat{W}^{t+1}$'s k-th column is either unchanged or decreased. By (91), we have

$$W_{ik}^{t+1} \le W_{ik}^{t}. \tag{93}$$

Since (93) is contradicted with (85), we know that $\left\{W^t\right\}$ is bounded. In the same way, we can prove that $\left\{H^{(u),t}\right\}$ is also bounded. Therefore, $\left\{H^t, W^t\right\}$ is in a compact set, and there is at least one convergence subsequence. This completes the proof.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[2] W. Wu, Y. Jia, S. Kwong, and J. Hou, "Pairwise constraint propagation-induced symmetric nonnegative matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6348–6361, Dec. 2018.

[3] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using L21-norm," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 673–682.

[4] Z. Li, J. Tang, and X. He, "Robust structured nonnegative matrix factorization for image representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1947–1960, May 2018.

[5] E. Y. Lam, "Non-negative matrix factorization for images with Laplacian noise," in *Proc. IEEE Asia Pacific Conf. Circuits Syst.*, Nov./Dec. 2008, pp. 798–801.

[6] A. B. Hamza and D. J. Brady, "Reconstruction of reflectance spectra using robust nonnegative matrix factorization," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3637–3642, Sep. 2006.

[7] N. K. Logothetis and D. L. Sheinberg, "Visual object recognition," *Annu. Rev. Neurosci.*, vol. 19, no. 1, pp. 577–621, Mar. 1996.

[8] S. E. Palmer, "Hierarchical structure in perceptual representation," *Cognit. Psychol.*, vol. 9, no. 4, pp. 441–474, Oct. 1977.

[9] E. Wachsmuth, M. W. Oram, and D. I. Perrett, "Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque," *Cerebral Cortex*, vol. 4, no. 5, pp. 509–522, 1994.

[10] L. Du, X. Li, and Y.-D. Shen, "Robust nonnegative matrix factorization via half-quadratic minimization," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 201–210.

[11] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Process. Lett.*, vol. 17, no. 1, pp. 4–7, Jan. 2010.

[12] H. Liu and Z. Wu, "Non-negative matrix factorization with constraints," in *Proc. 24th AAAI Conf. Artif. Intell.*, 2010, pp. 1–6.

[13] Y. Chen, M. Rege, M. Dong, and J. Hua, "Non-negative matrix factorization for semi-supervised data clustering," *Knowl. Inf. Syst.*, vol. 17, no. 3, pp. 355–379, 2008.

[14] Y. He, H. Lu, and S. Xie, "Semi-supervised non-negative matrix factorization for image clustering with graph Laplacian," *Multimedia Tools Appl.*, vol. 72, no. 2, pp. 1441–1463, 2014.

[15] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 1989.

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2001, pp. 1–7.

[17] D. Cai, X. He, X. Wu, and J. Han, "Non-negative matrix factorization on manifold," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 63–72.

[18] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang, "Constrained nonnegative matrix factorization for image representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1299–1311, Jul. 2012.

[19] C.-J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1589–1596, Nov. 2007.

[20] V. D. Blondel, N. D. Ho, and P. Dooren, "Weighted nonnegative matrix factorization and face feature extraction," *Image Vis. Comput.*, pp. 1–17, 2008.

[21] P. N. Belhumeour, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[22] D. B. Graham and N. M. Allinson, "Characterising virtual eigensignatures for general purpose face recognition," in *Face Recognition*. Berlin, Germany: Springer, 1998, pp. 446–456.

[23] W.-A. Lin, J.-C. Chen, C. D. Castillo, and R. Chellappa, "Deep density clustering of unconstrained faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8128–8137.

[24] H. Jiang, J. Jang, and O. Nachum, "Robustness guarantees for density clustering," *Proc. Mach. Learn. Res.*, vol. 89, pp. 3342–3351, Apr. 2019.

[25] S. A. Nene, S. K. Nayar, and H. Murase. (1996). *Columbia Object Image Library (Coil-100)*. [Online]. Available: https://www.kaggle.com/jessicali9530/coil100

[26] N. Guan, T. Liu, Y. Zhang, D. Tao, and L. S. Davis, "Truncated Cauchy non-negative matrix factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 246–259, Jan. 2019.

[27] R. Tanawongsuwan and A. Bobick, "Modelling the effects of walking speed on appearance-based gait recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. Jul. 2004, p. II.

[28] Z. Fu, K. Ren, J. Shu, X. Sun, and F. Huang, "Enabling personalized search over encrypted outsourced data with efficiency improvement," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 9, pp. 2546–2559, Sep. 2016.

[29] Y. Qin, S. Ding, L. Wang, and Y. Wang, "Research progress on semi-supervised clustering," *Cogn. Comput.*, pp. 1–14, 2019.

[30] D. Ienco and R. G. Pensa, "Semi-supervised clustering with multiresolution autoencoders," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2018, pp. 1–8.

[31] Z. Zhang, D. Kang, and C. Gao, "SemiSync: Semi-supervised clustering by synchronization," in *Proc. Int. Conf. Database Syst. Adv. Appl.* Cham, Switzerland: Springer, 2019, pp. 358–362.

[32] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in *Proc. IJCAI*, 2017, pp. 1753–1759.

[33] Y. Jia, S. Kwong, and J. Hou, "Semi-supervised spectral clustering with structured sparsity regularization," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 403–407, Mar. 2018.

[34] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, 2007.

[35] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.

[36] L. M. Abualigah, A. T. Khader, and E. S. Hanandeh, "A combination of objective functions and hybrid Krill herd algorithm for text document clustering analysis," *Eng. Appl. Artif. Intell.*, vol. 73, pp. 111–125, Aug. 2018.

[37] L. H. Son and T. M. Tuan, "Dental segmentation from X-ray images using semi-supervised fuzzy clustering with spatial constraints," *Eng. Appl. Artif. Intell.*, vol. 59, pp. 186–195, Mar. 2017.

[38] K. Zhao, W.-S. Chu, and A. M. Martinez, "Learning facial action units from Web images with scalable weakly supervised clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2090–2099.

[39] C. Li, Y. Tan, D. Wang, and P. Ma, "Research on 3D face recognition method in cloud environment based on semi supervised clustering algorithm," *Multimedia Tools Appl.*, vol. 76, no. 16, pp. 17055–17073, 2017.

[40] X. Yin, S. Chen, E. Hu, and D. Zhang, "Semi-supervised clustering with metric learning: An adaptive kernel method," *Pattern Recognit.*, vol. 43, no. 4, pp. 1320–1333, 2010.

[41] W. Kalintha, S. Ono, M. Numao, and K.-I. Fukui, "Kernelized evolutionary distance metric learning for semi-supervised clustering," in *Proc. 21st AAAI Conf. Artif. Intell.*, 2017, pp. 1–2.

[42] L. Lan, N. Guan, X. Zhang, D. Tao, and Z. Luo, "Soft-constrained non-negative matrix factorization via normalization," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 3025–3030.

[43] J. Zhao, X. Shang, and H. Zhang, "Recovering seabed topography from sonar image with constraint of sounding data," *J. China Univ. Mining Technol.*, vol. 46, no. 2, pp. 443–448, 2017.

[44] Y. Yan, L. Chen, and D. T. Nguyen, "Semi-supervised clustering with multi-viewpoint based similarity measure," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2012, pp. 1–8.

[45] X. Yin, T. Shu, and Q. Huang, "Semi-supervised fuzzy clustering with metric learning and entropy regularization," *Knowl.-Based Syst.*, vol. 35, pp. 304–311, Nov. 2012.

[46] X. Huang, X. Yang, J. Zhao, L. Xiong, and Y. Ye, "A new weighting k-means type clustering framework with an l2-norm regularization," *Knowl.-Based Syst.*, vol. 151, pp. 165–179, Jul. 2018.

[47] X. Liu, X. Zhu, M. Li, L. Wang, E. Zhu, T. Liu, M. Kloft, D. Shen, J. Yin, and W. Gao, "Multiple kernel k-means with incomplete kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[48] X. Shen, X. Zhang, L. Lan, Q. Liao, and Z. Luo, "Another robust NMF: Rethinking the hyperbolic tangent function and locality constraint," *IEEE Access*, vol. 7, pp. 31089–31102, 2019.

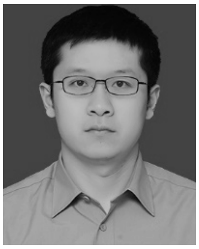
[49] S. Wei, Z. Li, and C. Zhang, "Combined constraint-based with metric-based in semi-supervised clustering ensemble," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 7, pp. 1085–1100, 2018.

LIUJING WANG received the B.S. degree in computer science from Shandong University, Jinan, China, in 2017. She is currently pursuing the master's degree with the National University of Defense Technology. Her current research interests include computer vision, matrix factorization, image processing, and object detection.

NAIYANG GUAN (M'10) received the B.S., M.S., and Ph.D. degrees from the National University of Defense Technology. He is currently an Associate Professor with the Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, China. He has authored or coauthored over 60 research articles on top-tier journals, including the IEEE T-PAMI, T-NNLS, T-IP, and T-SP, and top-tier conferences, including the IEEE ICDM, IJCAI, ECCV, and IJCNN. His research interests include machine learning, computer vision, and data mining.

DIANXI SHI received the B.S., M.S., and Ph.D. degrees in computer science from the National University of Defense Technology, Changsha, China, in 1989, 1996, and 2000, respectively. He is currently a Researcher and the Deputy Director of the Artificial Intelligence Research Center, National Innovation Institute of Defense Technology. His research interests include distributed object middleware technology, software component technology, adaptive software technology, and intelligent unmanned cluster system software architecture.

ZUNLIN FAN received the Ph.D. degree in electrical engineering from Air Force Engineering University. He is currently an Assistant Research Fellow with the Artificial Intelligence Research Center, National Innovation Institute of Defense Technology. His current research interests include statistical image processing, image denoising, image enhancement, and pattern recognition.



LONGFEI SU received the B.S. and M.S. degrees and the Ph.D. degree from the College of Automation and Mechanics, National University of Defense Technology. He is currently an Assistant Research Fellow. His current research interests include pattern recognition, statistical image processing, image denoising, and image enhancement.