Multi-View Subspace Clustering via Structured Multi-Pathway Network

Qianqian Wang, Zhiqiang Tao, Quanxue Gao, and Licheng Jiao, *Fellow, IEEE*

***Abstract*— Recently, deep multi-view clustering (MVC) has attracted increasing attention in multi-view learning owing to its promising performance. However, most existing deep multi-view methods use single-pathway neural networks to extract features of each view, which cannot explore comprehensive complementary information and multilevel features. To tackle this problem, we propose a deep structured multi-pathway network (SMpNet) for multi-view subspace clustering task in this brief. The proposed SMpNet leverages structured multi-pathway convolutional neural networks to explicitly learn the subspace representations of each view in a layer-wise way. By this means, both low-level and high-level structured features are integrated through a common connection matrix to explore the comprehensive complementary structure among multiple views. Moreover, we impose a low-rank constraint on the connection matrix to decrease the impact of noise and further highlight the consensus information of all the views. Experimental results on five public datasets show the effectiveness of the proposed SMpNet compared with several state-of-the-art deep MVC methods.**

***Index Terms*— Clustering structure, low-rank constraint, multipath network, multi-view clustering (MVC).**

I. Introduction

With the rapid development in data collection methods, massive data have been captured from different modalities, sensors, and sources, resulting in large-scale multi-view data that could describe the characteristics of the same instance from distinct perspectives [1]. However, most of the data are unlabeled, which brings in difficulty for data processing and analysis. One fundamental solution to this problem is to employ clustering analysis to group data into several clusters. Unfortunately, traditional clustering methods may not fully utilize the abundant information from different views. To explore the consensus and complementary information among multiple views, multi-view clustering (MVC) [2] has been proposed and has received considerable research attention in recent years. Existing works can be roughly divided into two categories, such as the traditional MVC methods [3], [4] and the deep MVC methods [5], [6].

On the one hand, the traditional MVC methods could be further divided into four directions: 1) the co-training-based methods alternately train between two views to maximize the mutual agreement; 2) the nonnegative matrix factorization (NMF)-based methods [7] recover a common indicator matrices from different views; 3) the graph-based methods [8], [9] try to construct a common similarity matrix across multiple views, which can further reveal data cluster structure; and 4) the subspace methods [10], [11] learn a common low-dimensional feature subspace shared by multiple views. All these above methods have achieved appealing performance for the clustering task. However, they may struggle to handle the data with high-dimensional features and nonlinear property, since they mainly adopt shallow and linear embedding functions to reveal the intrinsic structure of multi-view data.

On the other hand, deep neural networks are introduced to MVC for tackling the high-dimensional and nonlinear features, forming the research direction of deep MVC methods. Among previous works, deep multi-view subspace clustering (MVSC) methods [12], [13] have been well explored recently, which utilize a single-pathway network for learning an effective latent space from the original feature space [14], [15]. For example, Andrew et al. [16] proposed a deep canonical correlation analysis (DCCA) method to learn complex nonlinear transformations of two-view data such that the learned representations are highly linearly correlated. For another example, Abavisani et al. [17] employed convolutional neural networks for unsupervised multimodal subspace clustering (DMSC). Xie et al. [18] proposed a deep multi-view joint clustering (DMJC) in which multiple deep embedding features are learned simultaneously. However, although the deep MVSC methods have achieved promising results, there are still some challenges for MVSC. Specifically: 1) how to encode heterogeneous features, i.e. different level features, from multiple views with a minimum feature loss? 2) how to develop a deep network architecture to extract the view-consistent information, e.g., through the shared subspace representation, across different views? 3) how to explore the global structure of multi-view data?

To address the above challenges, we propose a novel deep structured multi-pathways network (SMpNet) for the multi-view subspace clustering as shown in Fig. 1. The proposed SMpNet consists of structured multi-pathway encoders/decoders, multilayer self-expression module, and low-rank subspace learning module. Inspired by multilevel feature extraction [19], [20], the structured multi-pathway encoders leverage multi-pathway convolutional neural networks to explicitly learn the subspace representation of each view in a layerwise way. By utilizing a connection matrix, the multilayer consistent self-expression layer integrates both high-level and low-level features into a common subspace. As we all known, in the CNN extraction, the low-level feature captures the detail features of images, such as line shape, etc. while the high-level feature is closer to semantic meanings. Thus, the obtained subspace could contain more comprehensive information than using a single-pathway to encode the multi-view data. Moreover, we impose a low-rank constraint on the consistent connection matrix against the impact of noise in multi-view data and to further enhance the cluster structure of the connection matrix, which is beneficial to the clustering task. The main contributions of this work are summarized as follows.

1) We construct a novel multi-pathway network for multi-view subspace clustering (SMpNet), which introduces multipath feature extraction instead of single-path way for each view.

Manuscript received 26 May 2021; revised 11 December 2021 and 28 June 2022; accepted 3 October 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62176203, in part by the Initiative Postdocs Supporting Program, in part by the China Postdoctoral Science Foundation under Grant 2019M653564, and in part by the Open Projects Program of National Laboratory of Pattern Recognition. *(Corresponding author: Quanxue Gao.)*

Qianqian Wang is with the State Key Laboratory of Integrated Services Networks, the Key Laboratory of Ministry of Education of Intelligence and Image Understanding, and the School of Telecommunications Engineering, Xidian University, Xi'an 710071, China.

Zhiqiang Tao is with the School of Information, Golisano College of Computing and Information Science, Rochester Institute of Technology, Rochester, NY 14623 USA (e-mail: zhiqiang.tao@rit.edu).

Quanxue Gao is with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: qxgao@xidian.edu.cn).

Licheng Jiao is with the Key Laboratory of Ministry of Education of Intelligence and Image Understanding, School of Artificial Intelligence, Xidian University, Xi'an 710071, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3213374.

Digital Object Identifier 10.1109/TNNLS.2022.3213374

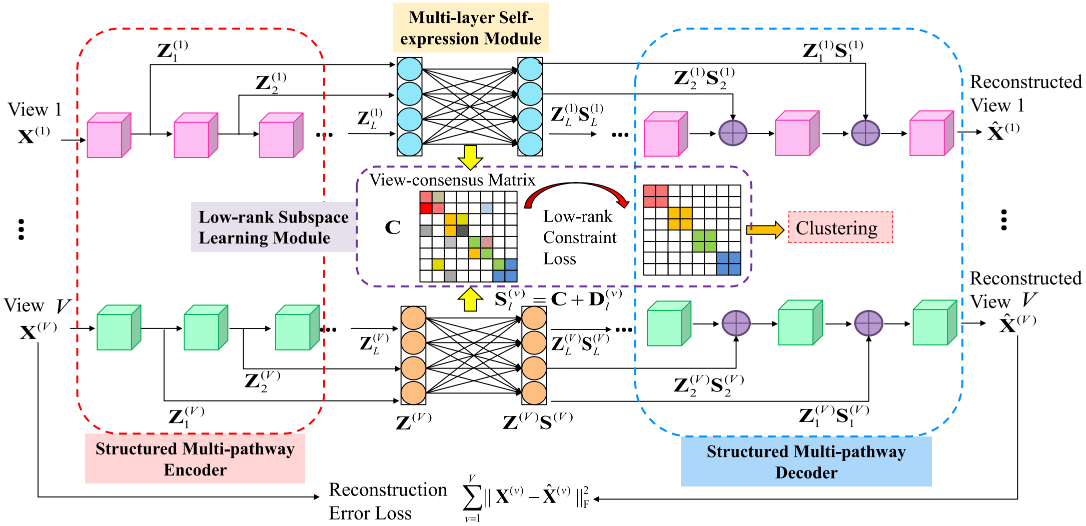
Multi-layer Self-expression Module
View 1
$\mathbf{X}^{(1)}$
$\mathbf{Z}_1^{(1)}$
$\mathbf{Z}_2^{(1)}$
$\mathbf{Z}_L^{(1)}$
$\mathbf{Z}_L^{(1)}\mathbf{S}_L^{(1)}$
$\mathbf{Z}_2^{(1)}\mathbf{S}_2^{(1)}$
$\mathbf{Z}_1^{(1)}\mathbf{S}_1^{(1)}$
Reconstructed View 1
$\hat{\mathbf{X}}^{(1)}$
View-consensus Matrix
C
Low-rank Constraint Loss
Low-rank Subspace Learning Module
Clustering
$\mathbf{S}_l^{(v)}=\mathbf{C}+\mathbf{D}_l^{(v)}$
View V
$\mathbf{X}^{(V)}$
$\mathbf{Z}_L^{(V)}$
$\mathbf{Z}_2^{(V)}$
$\mathbf{Z}_1^{(V)}$
$\mathbf{Z}^{(V)}$
$\mathbf{Z}^{(V)}\mathbf{S}^{(V)}$
$\mathbf{Z}_L^{(V)}\mathbf{S}_L^{(V)}$
$\mathbf{Z}_2^{(V)}\mathbf{S}_2^{(V)}$
$\mathbf{Z}_1^{(V)}\mathbf{S}_1^{(V)}$
Reconstructed View V
$\hat{\mathbf{X}}^{(V)}$
Structured Multi-pathway Encoder
Structured Multi-pathway Decoder
Reconstruction Error Loss
$\sum_{v=1}^{V}\|\mathbf{X}^{(v)}-\hat{\mathbf{X}}^{(v)}\|_F^2$

Fig. 1. Architecture of the proposed model deep SMpNet. SMpNet consists of a structured multi-pathway encoder, a structured multi-pathway decoder, a multilayer consistent self-expression module, and a low-rank subspace learning module. It aims to learn a better consistent connection matrix **C** for clustering.

By combining consensus-specific learning in the multi-pathway network, SMpNet could fully leverage both the high-level and low-level feature of each view and fuse more complementary information in a consistent subspace.

2) SMpNet adopts a low-rank constraint on the shared subspace to further explore the global structure feature among multiple views and remove noise and redundancy of the shared subspace. In this way, it ensures the obtained consensus representation achieves a better clustering for multi-view data.
3) We provide extensive experimental results on several public multi-view image datasets to show the effectiveness of the proposed SMpNet compared with the state-of-the-art methods. We also present a detailed ablation study for the proposed method.

II. Methodology

In this section, we first extended the low-rank subspace clustering model for single-view data to multi-view data and developed a deep low-rank multi-view subspace clustering model. Then, we described the structured multi-pathway network and applied it in deep low-rank MVC.

A. Deep Low-Rank Multi-View Subspace Clustering

Low-rank representation (LRR) aims at finding the lowest rank representation of all data jointly. Therefore, LRR is developed to uncover the global structures within the data and has been widely applied in clustering. For single-view data $\mathbf{X}$, the basic model of low-rank subspace clustering [21] is learning a LRR $\mathbf{S}$ by the following equation and conduct clustering on it:

$$\min_{\mathbf{S}} \|\mathbf{S}\|_* \quad \text{s.t. } \mathbf{X}=\mathbf{XS} \tag{1}$$

where $\|\cdot\|_*$ represents the nuclear norm, which is the sum of the singular values of the matrix and used to constrain the low rank of the matrix.

For deep multi-view subspace clustering networks, given N paired data samples $\{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(V)}\}_{i=1}^N$ from V different views, define the corresponding data matrices as $\mathbf{X}^{(v)} = [x_1^{(v)}, x_2^{(v)}, \ldots, x_N^{(v)}]$, $v \in \{1, 2, \ldots, V\}$. Regardless of the network structure, let $\Theta_e^{(v)}$ denote the parameters of the multi-view encoder. Similarly, let $\Theta_e^{(v)}$ be the self-expressive layer parameters and $\Theta_d^{(v)}$ be the multi-view decoder parameters. We aim to transform the original features with nonlinear mapping $f_v(\mathbf{X}^{(v)}; \Theta_e^{(v)}) \to \mathbf{Z}^{(v)}$. A self-expression layer is introduced between the deep encoder and the decoder to simulate the self-expression attributes and obtain consistent subspace descriptions $\mathbf{C} \in R^{N\times N}$ to improve clustering performance. In the self-expressive layer, we want $\mathbf{Z} = \mathbf{ZS}$ to implement self-expression properties. $\mathbf{Z}^{(v)}\mathbf{S}$ is input into vth deconvolutional decoders, which can generate all view reconstructed samples with the latent representations corresponding to each view. Specifically, $g_v(\mathbf{Z}^{(v)}\mathbf{S}; \Theta_d^{(v)}) \to \hat{\mathbf{X}}^{(v)}$, where $\hat{\mathbf{X}}^{(v)}$ represents the reconstructed sample matrix of the vth view. Then, the deep low-rank multi-view subspace clustering models can be trained end-to-end using the following loss function:

$$\begin{aligned}\min_{\Theta_e^{(v)},\ \Theta_d^{(v)},\mathbf{Z},\mathbf{S}} \ & \sum_{v=1}^{V}\left(\|\mathbf{X}^{(v)}-\hat{\mathbf{X}}^{(v)}\|_F^2+\alpha\|\mathbf{Z}^{(v)}-\mathbf{Z}^{(v)}\mathbf{S}\|_F^2\right)+\|\mathbf{S}\|_* \\ \text{s.t.} \ & \operatorname{diag}(\mathbf{S})=0.\end{aligned} \tag{2}$$

Finally, do spectral clustering (SC) on $\mathbf{S}$ to get the final clustering result. Compared with existing deep MVSC, (2) could explore global feature of multiple view by imposing low-rank constraint. However, (2) is still a single-pathway neural networks for each view as most deep MVC method, and hence they may not capture the rich feature context of each view. The inability to obtain multilevel features greatly limits the clustering performance of these methods, since they cannot take full advantage of the comprehensive feature embedded in multi-view data. To overcome this problem, we develop a structured multi-pathway CNN-based network, named as SMpNet, for the MVC task. It leverages multi-pathway CNNs to extract both high-level and low-level features of each single view. To further explore the view-consensus information, we separate the self-representation to view-consensus part and view-specific part, which helps to improve the clustering performance.

B. Structured Multi-Pathway Network

Fig. 1 illustrates the system structure of our proposed model SMpNet. It consists of a structured multi-pathway encoder, a multi-layer self-expression module, a low-rank subspace learning module, and a structured multi-pathway decoder. To encode heterogeneous features from multiple views and extract the view-consensus feature and view-specific feature of each view, SMpNet jointly learns the low-level and high-level feature of the input data by adding multiple convolution layers. They provide multiple paths of feature between the symmetrical layers of the encoders and the decoders. SMpNet can not only enhance the ability of the network in extracting more complex feature from the input data but also reconstruct the output of decoder layers to generate multiple sets of representations that satisfy the self-expressiveness property.

The overall objective function is composed of reconstruction error loss $\mathcal{L}_{\text{Re}}$, multilayer self-expression loss $\mathcal{L}_{\text{Se}}$, low-rank constraint loss $\mathcal{L}_C$, and regularize term $\mathcal{L}_D$ as follows:

$$\mathcal{L} = \mathcal{L}_{\text{Re}} + \lambda_1 \mathcal{L}_{\text{Se}} + \lambda_2 \mathcal{L}_C + \lambda_3 \mathcal{L}_D \tag{3}$$

where λ_1, λ_2, and λ_3 are tradeoff parameters that are set approximately to be inversely proportional to the value of each cost function.

1) Reconstruction Error Loss $\mathcal{L}_{\text{Re}}$: Since we have adopted an end-to-end training method to train the network, there is a reconstruction error loss between the original view $\mathbf{X}^{(v)}$, i.e. the input of structured multi-pathway encoder and the reconstructed view $\hat{\mathbf{X}}^{(v)}$, i.e. the output of structured multi-pathway decoder. Then, the reconstruction error loss of our network's end-to-end training is represented by the following:

$$\mathcal{L}_{\text{Re}} = \sum_{v=1}^{V} \left\| \mathbf{X}^{(v)} - \hat{\mathbf{X}}^{(v)} \right\|_F^2. \tag{4}$$

2) Multilayer Self-Expression Loss $\mathcal{L}_{\text{Se}}$: After the structured multi-pathway encoder, we obtain L latent subspaces $\mathbf{Z}_l^{(v)}$ for each view, where $\mathbf{Z}_l^{(v)}$ denotes the learned latent subspace from the lth layer convolutional network in the vth view. Then, using multilayer self-expression module, we learn self-expression matrix $\mathbf{S}_l^{(v)}$ by $\mathbf{Z}_l^{(v)} = \mathbf{Z}_l^{(v)}\mathbf{S}_l^{(v)}$. Considering that the multiple view features cover the information of the same data, thus there should exist common information shared among different views. We separate self-expression matrix $\mathbf{S}_l^{(v)}$ with the view-consensus matrix $\mathbf{C}$ of multiple views and view-specific matrix $\mathbf{D}_l^{(v)}$ of the lth layer in vth view. simultaneously. The loss function of the self-expression for the proposed SMpNet is computed by

$$\begin{aligned} \mathcal{L}_{\text{Se}} &= \sum_{l=1}^{L} \sum_{v=1}^{V} \left\| \mathbf{Z}_l^{(v)} - \mathbf{Z}_l^{(v)}\mathbf{S}_l^{(v)} \right\|_F^2 \\ \text{s.t. } \mathbf{S}_l^{(v)} &= \mathbf{C} + \mathbf{D}_l^{(v)}, \quad \text{diag}(\mathbf{C}) = 0 \end{aligned} \tag{5}$$

where $\mathbf{C}$ describes the common feature subspace, and $\mathbf{D}_l^{(v)}$ captures the unique information of each individual layer across views. To avoid the trivial solution $\mathbf{C} = \mathbf{I}$, we constrain $\text{diag}(\mathbf{C}) = 0$.

3) Low-Rank Constraint Loss $\mathcal{L}_C$: For data with sparse properties, it should exhibit to be low rank and contain redundant information, which can be used to recover data and extract features. In addition, to obtain more global feature of multi-view data, we impose a low-rank constraint on view-consensus matrix $\mathbf{C}$ to excavate the more consistent information among different views by low-rank subspace learning module. The low-rank constraint loss is defined as follows:

$$\mathcal{L}_C = \|\mathbf{C}\|_* \tag{6}$$

where $\|\cdot\|_*$ represents the nuclear norm, which is the sum of the singular values of the matrix and used to constrain the low rank of the matrix.

Meanwhile, there is some unique information in each view that could compensate for the accurate reconstruction. The l_2 regularization tends to be smaller and more diffuse weight vectors, which encourages the diversity matrices to eventually use features on all dimensions rather than relying heavily on a few of them. Hence, we further define a regularization loss to guarantee a nontrivial solution of the diverse coefficient matrix as

$$\mathcal{L}_D = \sum_{l=1}^{L} \sum_{v=1}^{V} \left\| \mathbf{D}_l^{(v)} \right\|_F^2. \tag{7}$$

Using $\mathcal{L}_D$, the generalization ability of our model can be improved and the risk of overfitting can be also reduced.

Algorithm 1 Structured Multi-Pathway Network for Multi-View Subspace Clustering

```
Input:
    Multi-view dataset $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$, $\lambda_1$, $\lambda_2$, $\lambda_3$.
Initial:
    Learning rate: lr=0.001, Epoch=5000.
    Whlie pretraining not converged do:
        Update encoder $\Theta_e^{(v)}$, decoder $\Theta_d^{(v)}$, $\mathbf{C}$, and $\mathbf{D}^{(v)}$ by minimizing Eq. (3).
    End pretraining.
    While training not converged do:
        Update encoder $\Theta_e^{(v)}$ and decoder $\Theta_d^{(v)}$, $\mathbf{C}$, and $\mathbf{D}^{(v)}$ by minimizing Eq. (3).
    End training
    return  C and $\mathbf{D}^{(v)}$.
    Use $\mathbf{W} = 1/2(|\mathbf{C}|^\top + |\mathbf{C}|)$ as the affinity matrix of the spectral cluster to do MVC.
    End Procedure
```

C. Training Procedure

In our experiment, we iteratively optimize the parameters of the structured multi-pathway encoder, structured multi-pathway decoder, and self-expression layer to get our final parameters. First, the multi-view data $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$ are input into the encoder network to obtain the hidden layer features $\mathbf{Z}_l^{(v)}$ and the self-expression feature $\mathbf{Z}_l^{(v)}\mathbf{S}_l^{(v)}$. Then, the decoder obtains reconstructed multi-view data $\hat{\mathbf{X}}^{(v)}$ with $\mathbf{Z}_l^{(v)}\mathbf{S}_l^{(v)}$ as input. Finally, $\Theta_e^{(v)}$, $\Theta_d^{(v)}$, $\mathbf{C}$, and $\mathbf{D}_l^{(v)}$ are optimized by minimizing the loss function of the overall network.

After obtaining the consistent connection matrix $\mathbf{C}$, we normalize each column of it to get normalized $\mathbf{W}$ by $\mathbf{W} = 1/2(|\mathbf{C}|^\top + |\mathbf{C}|)$. $\mathbf{W}$ can work as the affinity matrix between the clustering samples to perform SC for the final clustering results. The detailed training procedures are summarized in Algorithm 1.

III. Experiments

A. Experimental Settings

We compare our approach with the contrast algorithms with the five datasets including Fashion-MNIST dataset, COIL20 dataset, MNIST dataset, YoutubeFace (YTF) dataset, and FRGC dataset. Clustering accuracy (ACC) [22] and normalized mutual information (NMI) [23] are used to compare the performance of different methods.

1) Datasets:

a) Fashion-MNIST dataset [24]: The size, format, and training set/test set partition of Fashion-MNIST is exactly the same as the original MNIST. In our experiment, we randomly choose 200 images of each category and disordered them.

b) COIL-20 dataset [25]: COIL-20 dataset collected 1440 grayscale image data of 20 categories from different shooting angles.

c) FRGC dataset [26]: Using the 20 random selected subjects in [26] from the original dataset, we collect 2462 face images.

TABLE I

STATISTICS OF MULTIMODAL REAL-WORLD DATASETS. NOTE THAT THE TRAINING AND TESTING IMAGES IN EACH DATASET ARE JOINTLY UTILIZED FOR CLUSTERING

Dataset	#Sample number	#Class number	#View number
Fashion MNIST	70000	10	2
COIL-20	1440	20	2
YTF	10000	41	3
FRGC	2462	20	3
MNIST	70000	10	2

d) YoutubeFace (YTF) dataset: [26] Similar to FRGC dataset, we selected the first 41 topics (2000 images) of the YTF dataset.

e) MNIST dataset [27]: consists of 70 000 handwritten digit images with 28 × 28 pixels and is a widely-used benchmark dataset. In our experiment, we choose 200 images of each class in MNIST. Detailed information of these datasets in our experiment has been shown in Table I. For Fashion-MNIST, and MNIST dataset, we use grayscale images, edge feature images as two views. As for COIL-20 dataset, considering that the database dimension is too large, we only use two views here. The first view is the original images, and the second view is the edge feature images. For the two color image datasets (YTF dataset and FRGC dataset), we use the original RGB images, grayscale images, and edge feature images as three views.

2) Comparison Algorithms: We choose several state-of-the-art MVC compared algorithms as baseline.

1) *Single-View Clustering Methods*: K-means clustering [28], SC [29], deep embedded clustering (DEC) [30], and adaptive self-paced deep clustering with data augmentation (ASPC-DA) [31].
2) *Traditional MVC Methods*: Robust multi-view K-means clustering (RMKMC) [32], binary MVC (BMVC) [33], graph-based MVC (GMC) [34], and locality adaptive latent MVC (LALMVC) [35].
3) *Deep MVC Methods*: DCCA [16], deep typical correlated auto-encoder (DCCAE) [36], deep generalized canonical correlation analysis (DGCCA) [37], joint framework for DMJC [38], deep multi-view subspace clustering (DMSC) [17], and cross-modal subspace clustering via DCCA (CMSC-DCCA) [14].

3) Implementation Details: For each convolutional encoder, we set up a three-layer network. The size of the first-layer convolution kernel is 4 × 4 × 10 and the step size is 2. The size of the second-layer convolution kernel is 3 × 3 × 20 and the step size is 1. The third layer convolution kernel size is 4 × 4 × 30 and the step size is 2. The deconvolution decoder has a convolution kernel size opposing to that of the convolution encoder. Specifically, for the YTF dataset, the second layer output is zero-padded to match the dimension. We use PyTorch's public toolbox to implement our method and other nonlinear methods. All experiments are run on the Ubuntu Linux 16.04 platform with NVIDIA Titan Xp graphics processing unit (GPU) and 64 GB of memory. The Adam [39] optimizer with default parameter settings is used to train our model and fix the learning rate to 0.001. In the training procedure, 5000 optimization epochs are performed. The batch size in the optimization process is set as the data size of the input data, so that the subspace clustering is consistent in dimension. All other linear methods are tested in the same environment as MATLAB. Our source code has been uploaded to Github website: https://github.com/IMKBLE/SMpNet.

B. Experiments Analysis

1) Clustering Performance With Comparison Algorithms: The ACC and NMI results on all databases are shown in Table II, where the best results in each dataset are highlighted in **bold**. For single-view clustering methods, we only give the best single-view results in the table. From Table II, we have the several observations as follows: 1) compared with single-view methods (Kmeans, SC, DEC, ASDC-DA), we can see that almost all the MVC methods achieve better ACC and NMI than all best single-view clustering methods. It illustrates that multi-view methods are more effective than single-view methods when processing multi-view data. It further indicates that multi-view data provide more abundant information. On MNIST dataset, one single-view deep clustering methods ASDC-DA is superior to others. It probably because ASDC-DA uses data augmentation and self-supervised learning to boost the clustering performance; 2) most deep MVC methods (DGCCA, DMJC, DMSC, CMSC-DCCA) are superior to traditional MVC methods. It is because deep MVC methods utilize the neural network to learn the nonlinear feature in multi-view data, which contains more discriminative information; and 3) the proposed SMpNet almost achieves the best ACC and NMI in most datasets among the multi-view methods. This is mainly because other deep multi-view methods only use single-pathway neural networks. Differently, SMpNet leverages multi-pathway convolutional neural networks to extract high-level and low-level features of each view, which will be further integrated into a common subspace via the consistent self-expression layer. Besides, SMpNet imposes a low-rank constraint on the shared subspace representations to explore the consistent information among multiple views, which further improves the clustering performance.

2) Ablation Study: We further conduct ablation studies for SMpNet to verify how much each term in the objective function contributes to the clustering performance. First, to verify the performance of multi-pathway structure, we add two more baseline by running SC on the best single Z^v given by SMpNet, and running SC on SMpNet with single pathway. Table III shows the results. From the results, we can see the proposed SMpNet with multiple pathways is superior to SMpNet with a single pathway.

Then, we perform experiments in five cases to isolate the effect of the reconstruction error loss $\mathcal{L}_{\text{Re}}$, multilayer self-expression loss $\mathcal{L}_{\text{Se}}$, regularization constraint $\mathcal{L}_D$, and low-rank constraint $\mathcal{L}_C$: 1) SMpNet w/o $\mathcal{L}_{\text{Re}}$ (where w/o represents without.); 2) SMpNet w/o $\mathcal{L}_{\text{Se}}$; 3) SMpNet w/o $\mathcal{L}_D$; 4) SMpNet w/o $\mathcal{L}_C$; and 5) SMpNet with overall objective function. Table III presents the experimental results, and we could see our MMSC achieves the best performance when using the overall objective, which demonstrates the necessity of the four losses. Besides, SMpNet w/o $\mathcal{L}_{\text{Re}}$ achieves the worst ACC and NMI on all the four databases, which means the reconstruction error loss contributes most to the clustering performance. In the case, SMpNet w/o $\mathcal{L}_D$, SMpNet without the regularization constraint $\mathcal{L}_D$ shows similar ACC and NMI with MMSC with the overall objective, which illustrates regularization constraint has the smallest effect on the SMpNet model. The contribution of multilayer self-expression loss is between reconstruction loss and regularization constrain/low-rank constraint.

To show the results visually, we visualize the consistent connection matrixes $\mathbf{C}$ which are learned by SMpNet w/o $\mathcal{L}_{\text{Re}}$, SMpNet w/o $\mathcal{L}_C$, SMpNet w/o $\mathcal{L}_{\text{Se}}$ and SMpNet with overall objective function on FRGC dataset. Fig. 2 respectively shows the visualized results. The value of points within each diagonal block in the visualization of the consistent connection matrixes $\mathbf{C}$ denote the affinity of two corresponding samples in one cluster. Hence, if the diagonal block

TABLE II

OPTIMAL CLUSTERING ACCURACY (ACC %) AND THE NORMALIZED MUTUAL INFORMATION (NMI %) ON ALL THE DATASETS. (BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST RESULTS ARE UNDERLINED)

Methods	Fashion-MNIST		COIL-20		FRGC		YTF		MNIST	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Best K-means [28]	51.27	49.99	57.49	73.22	23.62	27.12	56.01	75.23	50.82	52.59
Best SC [29]	50.75	49.03	45.76	68.67	40.74	47.42	56.80	74.90	54.17	47.46
Best DEC [30]	51.80	54.60	68.00	80.25	37.80	50.50	37.10	44.60	65.70	59.89
Best ASPC-DA [31]	58.45	58.70	61.04	67.90	33.51	32.27	54.95	71.68	**73.95**	**78.65**
RMKMC [32]	53.32	52.87	60.97	74.93	23.52	25.85	57.21	74.56	52.40	50.44
BMVC [33]	45.36	38.05	34.31	40.33	41.51	45.92	28.13	38.28	20.65	18.02
GMC [34]	56.70	62.90	74.17	82.50	56.09	64.18	55.40	74.22	44.65	58.70
LALMVC [35]	52.15	59.96	74.31	83.77	52.56	65.73	55.60	71.51	53.85	58.42
DCCA [16]	52.74	53.82	55.76	64.91	22.91	24.75	45.19	60.35	48.15	46.87
DCCAE [36]	51.87	53.01	61.60	71.56	32.33	31.22	45.57	60.15	51.95	41.28
DGCCA [37]	56.28	57.04	54.01	62.40	23.76	24.53	47.26	61.38	47.85	46.68
DMJC [38]	61.41	63.41	72.99	81.58	44.07	59.79	61.15	77.40	53.55	45.43
DMSC [17]	59.55	65.07	74.10	86.82	72.83	80.96	62.80	80.16	67.75	65.00
CMSC-DCCA [14]	**62.95**	**68.33**	82.64	91.45	70.80	78.55	66.15	82.67	69.90	68.87
SMpNet	62.35	64.76	**85.69**	**93.23**	**76.12**	**81.28**	**68.60**	**84.27**	70.40	69.21

TABLE III

ABLATION STUDY OF THE PROPOSED METHOD SMPNET

Methods	Fashion-MNIST		COIL-20		FRGC		YTF		MNIST	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Best single Z^v by SMpNet	60.05	62.25	81.06	88.65	65.68	73.16	62.30	76.35	63.30	65.53
SMpNet with single pathway	59.50	59.72	82.15	91.32	67.18	79.13	44.50	60.02	57.50	60.54
SMpNet w/o $\mathcal{L}_{Re}$	44.95	46.37	51.92	54.09	48.84	50.61	47.43	58.51	53.65	60.33
SMpNet w/o $\mathcal{L}_{Se}$	49.23	52.89	68.27	67.85	62.10	65.47	57.38	62.19	64.90	62.76
SMpNet w/o $\mathcal{L}_D$	60.51	62.63	84.08	89.31	73.28	77.86	66.39	82.63	65.05	62.91
SMpNet w/o $\mathcal{L}_C$	59.97	60.18	83.90	88.75	74.29	76.69	65.76	81.78	67.60	66.29
SMpNet	**62.35**	**64.76**	**85.69**	**93.23**	**76.12**	**81.28**	**68.60**	**84.27**	**70.40**	**69.21**

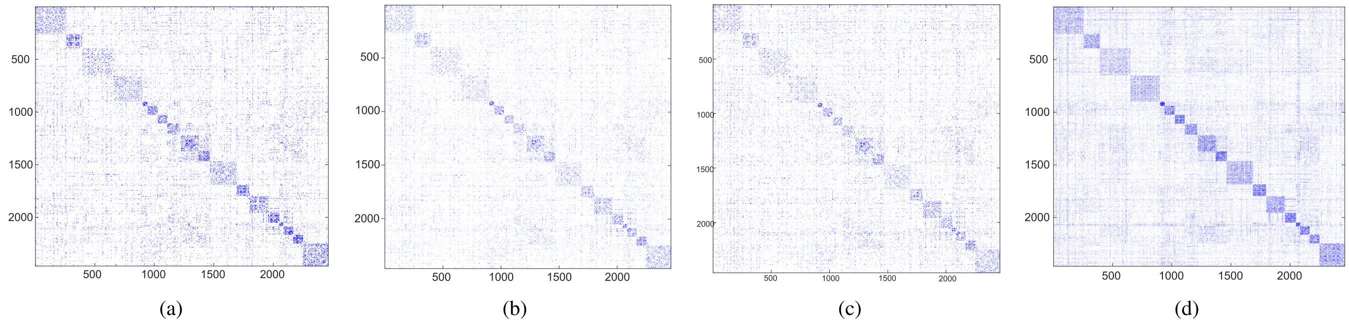

Fig. 2. Visualization of connection matrix $\mathbf{C}$ which are learned by the proposed method with different objective function on FRGC. (a) MMSC w/o $\mathcal{L}_{\text{Re}}$. (b) MMSC w/o $\mathcal{L}_C$. (c) MMSC w/o $\mathcal{L}_{\text{Se}}$. (d) MMSC (full model).

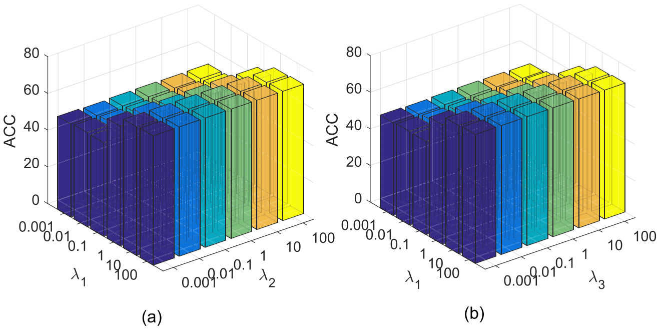

Fig. 3. Impact of parameters on clustering performance on MNIST dataset. (a) λ_3 = 1. (b) λ_2 =1.

of a similarity matrix is clearer and has more nonzero points, the clustering methods will perform better. From Fig. 2, we can see SMpNet with overall objective has a clearer diagonal block. Thus, it has the best clustering performance. The conclusion is similar with Table III.

3) Parameter Analysis: In our model, there are three parameters λ_1, λ_2, and λ_3. We use the method of controlling variables to analyze the parameters. Firstly, we fix the parameter λ_1 of the multilayer self-expression loss $\mathcal{L}_{\text{Se}}$ and vary both the parameters λ_2 and λ_3 of the regularize term $\mathcal{L}_D$ and low-rank constraint loss $\mathcal{L}_C$ in the range of $\{0.001, 0.01, 0.1, 1, 10, 100\}$. Then, we fix λ_2 and vary λ_1 and λ_3. Since the strategies of setting parameters are the same on all the four datasets, we only show the effect of parameters on MNIST dataset. From Fig. 3, we can notice that 1) our method can achieve the best

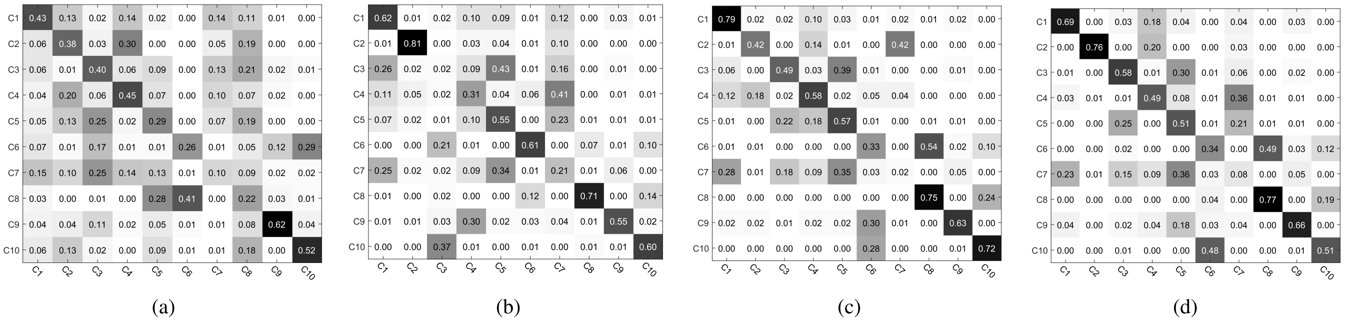

Fig. 4. Visualization of the confusion matrix for consistent connection matrix **C** in different epoch. (a) Epoch = 100. (b) Epoch = 1000. (c) Epoch = 1500. (d) Epoch = 2000.

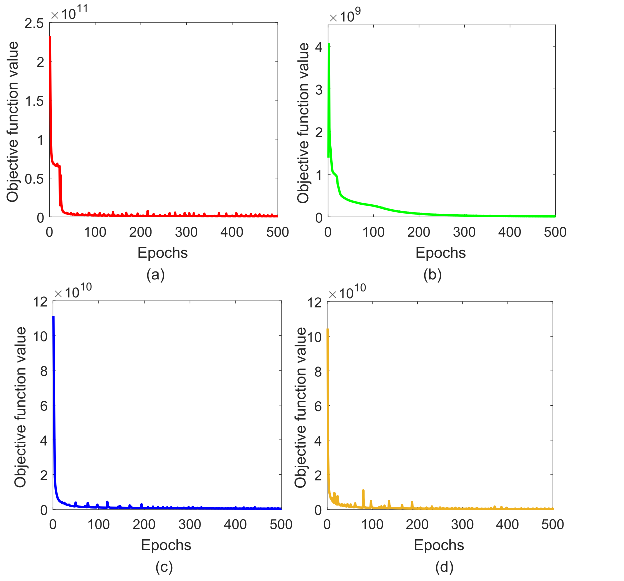

Fig. 5. Convergence curves of the proposed SMpNet on four datasets. (a) COIL-20. (b) Fashion-MNIST. (c) FRGC. (d) YTF.

ACC results on MNIST dataset when $\lambda_1 = 1$, $\lambda_2 = 0.1$ and $\lambda_3 = 1$; 2) our method is stable since varying parameters has little influence on the clustering performance. For Fashion-MNIST, COIL-20, and FRGC datasets, we set λ_1, λ_2, and λ_3 as 0.1, 1, and 0.01, respectively; For YTF dataset, we set λ_1, λ_2, and λ_3 as 0.01, 0.1, and 0.001, respectively.

4) Convergence Analysis: To investigate the convergence of our model, we visualize the confusion matrix of consistent connection matrix **C** which are learned by our model in different epochs on FRGC. In the consistent connection confusion matrix representation diagram, the diagonal elements are the correct clustering rate for each type of samples. The higher the value, the deeper the corresponding color, and the better the clustering results of the reaction. Fig. 4 shows the results of the confusion matrix. We can see, with the increase of epoch, the color of the diagonal elements in the corresponding confusion matrix becomes deeper. Thus, the clustering performance of the proposed methods becomes better. In addition, in the experiments, we record the objective function value of the proposed model for each ten epochs. Fig. 5 shows the objective function value with increasing epochs on the four databases. As seen, the proposed model converges very quickly when the epochs are less than 500 times. This result ensures the speed of the whole proposed method.

IV. Conclusion

In this brief, we propose a deep multi-view subspace clustering with structured multi-pathway neural network. In the proposed SMpNet, it jointly learns the low-level and high-level features of the input data from the multipath convolutional networks. Thus, the obtained common subspace contains more complex feature of input multi-view data than other DMVC methods. Besides, SMpNet integrates low-rank constrain on the consistent connection matrix to remove the influence of noise. We validate the clustering performance improvement of SMpNet via a series of comprehensive experiments, and the comparison results to several existing methods demonstrate the superiority of SMpNet.

References

[1] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," 2013, *arXiv:1304.5634*.

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

[3] X. Liu, Y. Dou, J. Yin, L. Wang, and E. Zhu, "Multiple kernel k-means clustering with matrix-induced regularization," in *Proc. AAAI*, 2016, pp. 1888–1894.

[4] X. Zhu, S. Zhang, W. He, R. Hu, C. Lei, and P. Zhu, "One-step multi-view spectral clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, pp. 2022–2034, Oct. 2019.

[5] R. Zhou and Y.-D. Shen, "End-to-end adversarial-attention network for multi-modal clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14619–14628.

[6] C. Zhang et al., "Generalized latent multi-view subspace clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 86–99, Jan. 2020.

[7] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. SIAM Int. Conf. Data Mining*, May 2013, pp. 252–260.

[8] F. Nie, G. Cai, and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours," in *Proc. AAAI*, 2017, pp. 2408–2414.

[9] Z. Tao, H. Liu, S. Li, Z. Ding, and Y. Fu, "Marginalized multiview ensemble clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 2, pp. 600–611, Apr. 2019.

[10] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[11] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[12] Z. Huang, J. T. Zhou, H. Zhu, C. Zhang, J. Lv, and X. Peng, "Deep spectral representation learning from multi-view data," *IEEE Trans. Image Process.*, vol. 30, pp. 5352–5362, 2021.

[13] S. Chang, J. Hu, T. Li, H. Wang, and B. Peng, "Multi-view clustering via deep concept factorization," *Knowl.-Based Syst.*, vol. 217, Apr. 2021, Art. no. 106807.

[14] Q. Gao, H. Lian, Q. Wang, and G. Sun, "Cross-modal subspace clustering via deep canonical correlation analysis," in *Proc. AAAI*, 2020, pp. 3938–3945.

[15] Q. Wang, W. Xia, Z. Tao, Q. Gao, and X. Cao, "Deep self-supervised t-SNE for multi-modal subspace clustering," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1748–1755.

[16] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. ICML*, 2013, pp. 1247–1255.

[17] M. Abavisani and V. M. Patel, "Deep multimodal subspace clustering networks," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 6, pp. 1601–1614, Dec. 2018.

[18] Y. Xie et al., "Joint deep multi-view learning for image clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 11, pp. 3594–3606, Nov. 2021.

[19] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. ECCV*, 2018, pp. 527–542.

[20] M. Kheirandishfard, F. Zohrizadeh, and F. Kamangar, "Multi-level representation learning for deep subspace clustering," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2039–2048.

[21] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[22] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[23] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.

[24] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.

[25] F. Li, H. Qiao, and B. Zhang, "Discriminatively boosted image clustering with fully convolutional auto-encoders," *Pattern Recognit.*, vol. 83, pp. 161–173, Nov. 2018.

[26] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5147–5156.

[27] C. Shang, A. Palmer, J. Sun, K.-S. Chen, J. Lu, and J. Bi, "VIGAN: Missing view imputation with generative adversarial networks," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 766–775.

[28] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.

[29] A. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. NeurIPS*, 2001, pp. 849–856.

[30] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. ICML*, 2016, pp. 478–487.

[31] X. Guo et al., "Adaptive self-paced deep clustering with data augmentation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 9, pp. 1680–1693, Sep. 2020.

[32] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *Proc. IJCAI*, 2013, pp. 2598–2604.

[33] Z. Zhang, L. Liu, F. Shen, H. T. Shen, and L. Shao, "Binary multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1774–1782, Jul. 2018.

[34] H. Wang, Y. Yang, and B. Liu, "GMC: Graph-based multi-view clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1116–1129, May 2019.

[35] D. Xie, X. Zhang, Q. Gao, J. Han, S. Xiao, and X. Gao, "Multiview clustering by joint latent representation and similarity learning," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4848–4854, Nov. 2020.

[36] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning: Objectives and optimization," 2016, *arXiv:1602.01024*.

[37] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, "Deep generalized canonical correlation analysis," 2017, *arXiv:1702.02519*.

[38] B. Lin, Y. Xie, Y. Qu, C. Li, and X. Liang, "Jointly deep multi-view learning for clustering analysis," 2018, *arXiv:1808.06220*.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.