Multi-Level Representation Learning for Deep Subspace Clustering

Mohsen Kheirandishfard Fariba Zohrizadeh Farhad Kamangar
Department of Computer Science and Engineering,
University of Texas at Arlington, USA
{mohsen.kheirandishfard,fariba.zohrizadeh}@gmail.com kamangar@cse.uta.edu

Abstract

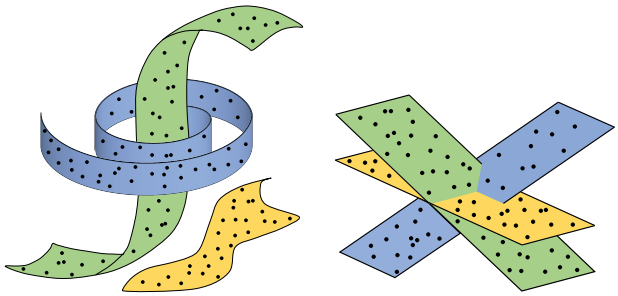
This paper proposes a novel deep subspace clustering approach which uses convolutional autoencoders to transform input images into new representations lying on a union of linear subspaces. The first contribution of our work is to insert multiple fully-connected linear layers between the encoder layers and their corresponding decoder layers to promote learning more favorable representations for subspace clustering. These connection layers facilitate the feature learning procedure by combining low-level and high-level information for generating multiple sets of self-expressive and informative representations at different levels of the encoder. Moreover, we introduce a novel loss minimization problem which leverages an initial clustering of the samples to effectively fuse the multi-level representations and recover the underlying subspaces more accurately. The loss function is then minimized through an iterative scheme which alternatively updates the network parameters and produces new clusterings of the samples. Experiments on four real-world datasets demonstrate that our approach exhibits superior performance compared to the state-of-the-art methods on most of the subspace clustering problems.

1. Introduction

Subspace clustering is an unsupervised learning task with a variety of machine learning applications such as motion segmentation [20, 36], face clustering [3, 51], movie recommendation [27, 50], etc. The primary goal of this task is to partition a set of data samples, drawn from a union of low-dimensional subspaces, into disjoint clusters such that the samples within each cluster belong to the same subspace [2, 28]. A large body of subspace clustering literature relies on the concept of self-expressiveness which states that each sample point in a union of subspaces is efficiently expressible in terms of a linear (or affine) combination of other points in the subspaces [8]. Given that, it is expected that the nonzero coefficients in the linear representation of each sample correspond to the points of the same subspace as the given sample. In order to successfully infer such underlying relationships among the samples and to partition them into their respective subspaces, a common practice approach is to first learn an affinity matrix from the input data and then apply the spectral clustering technique [26] to recover the clusters. Recently, these spectral clustering-based approaches have shown special interest in utilizing sparse or low-rank representations of the samples to create more accurate affinity matrices [8, 9, 22, 23, 41]. A well-established instance is sparse subspace clustering (SSC) [8] which uses an ℓ_1-regularized model to select only a small subset of points belonging to the same subspace for reconstructing each data point. More theoretical and practical aspects of the SSC algorithm are investigated and studied in detail in [34, 38, 48, 49].

Despite the key role that the self-expressiveness plays in the literature, it may not be satisfied in a wide range of applications in which samples lie on non-linear subspaces, e.g. face images taken under non-uniform illumination and at different poses [16]. A common practice technique to handle these cases is to leverage well-known kernel trick to implicitly map the samples into a higher dimensional space so that they better conform to linear subspaces [29, 30, 42, 46]. Although this strategy has demonstrated empirical success, it is not widely applicable to various applications, mainly because identifying an appropriate kernel function for a given set of data points is a quite difficult task [54].

Recently, deep neural networks have exhibited exceptional ability in capturing complex underlying structures of data and learning discriminative features for clustering [6, 11, 17, 40, 43]. Inspired by that, a new line of research has been established to bridge deep learning and subspace clustering for developing deep subspace clustering approaches [1, 16, 33, 44, 55]. Variational Autoencoders (VAE) [19, 24] and Generative Adversarial Network (GAN) [12] are among the most popular deep architectures adopted by these methods to produce feature representations suitable for subspace clustering [24]. Compared to the conventional approaches, deep subspace clustering methods



(a) (b)

Figure 1: Illustration of representation learning for subspace clustering. (a) Sample points may come from a union of nonlinear subspaces; (b) Deep subspace clustering approaches aim to transform the samples into a latent space so that they lie in a union of linear subspaces.

can better exploit the non-linear relationships between the sample points and consequently they achieve superior performance, especially in complex applications in which the samples do not necessarily satisfy the self-expressiveness property [16].

In this paper, we propose a novel spectral clustering-based approach which utilizes stacked convolutional autoencoders to tackle the problem of subspace clustering. Inspired by the idea of residual networks, our first contribution is to add multiple fully-connected linear layers between the corresponding layers of the encoder and decoder to infer multi-level representations from the output of every encoder layer. These connection layers enable to produce representations which are enforced to satisfy self-expressiveness property and hence well-suited to subspace clustering. We model each connection layer as a self-expression matrix created from the summation of a coefficient matrix shared between all layers and a layer-specific matrix that captures the unique knowledge of each individual layer. Moreover, we introduce a novel loss function that utilizes an initial clustering of the samples and efficiently aggregates the information at different levels to infer the coefficient matrix and the layer-specific matrices more accurately. This loss function is further minimized in an iterative scheme which alternatively updates the network parameters for learning better subspace clustering representations and produces a new clustering of the samples. We perform extensive experiments on four benchmark datasets for subspace clustering, including two face image and two object image datasets, to evaluate the efficacy of the proposed method. The experiments demonstrate that our approach can efficiently handle clustering the data from non-linear subspaces and it performs better than the state-of-the-art methods on most of the subspace clustering problems.

2. Related Works

Conventional subspace clustering approaches aim to learn a weighted graph whose edge weights represent the relationships between the samples of input data. Then, spectral clustering [26] (or its variants [37]) can be employed to partition the graph into a set of disjoint sub-graphs corresponding to different clusters [4, 7, 8, 9, 13, 15, 22, 35, 47]. A commonly-used formulation to obtain such a weighted graph is written as

$$\underset{\mathbf{C}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{C}\|_{\mathrm{F}}^{2}+\lambda\, g(\mathbf{C}) \tag{1a}$$

$$\text{subject to} \quad \operatorname{diag}(\mathbf{C})=\mathbf{0}, \tag{1b}$$

where $\|.\|_{\mathrm{F}}$ indicates Frobenius norm, $\mathbf{X}\in\mathbb{R}^{d\times n}$ is a data matrix with its columns representing the samples $\{\mathbf{x}_i\in\mathbb{R}^d\}_{i=1}^n$, $\mathbf{C}$ is a self-expression matrix with its $(i,j)^{th}$ element denoting the contribution of sample $\mathbf{x}_j$ in reconstructing $\mathbf{x}_i$, $g:\mathbb{R}^{n\times n}\rightarrow\mathbb{R}$ is a certain regularization function, and $\lambda>0$ is a hyperparameter to balance the importance of the terms. Equality constraint (1b) is imposed to eliminate the trivial solution $\mathbf{C}=\mathbf{I}_n$ that represents a point as a linear combination of itself. Once the optimal solution $\overset{*}{\mathbf{C}}$ of (1a) – (1b) is obtained, symmetric matrix $\frac{1}{2}(|\overset{*}{\mathbf{C}}|+|\overset{*}{\mathbf{C}}|^{\top})$ can serve as the affinity matrix of the desired graph where $|.|$ shows the element-wise absolute value operator. Different variants of (1a) – (1b) have been well-studied in the literature where they utilize various choices of the regularization function $g(.)$ such as $\|\mathbf{C}\|_0$ [45, 48], $\|\mathbf{C}\|_1$ [8], $\|\mathbf{C}\|_*$ [23, 41], $\|\mathbf{C}\|_{\mathrm{F}}$ [34], etc, to impose desired structures on the graph.

Deep generative architectures, most notably GANs and VAEs, have been widely used in the recent literature to facilitate the clustering task [25], especially when the samples come from complex and irregular distributions [24, 43]. These architectures improve upon the conventional feature extractions by learning more informative and discriminative representations that are highly suitable for clustering [5, 32, 33]. To promote inferring clusters with higher quality, some deep approaches propose to jointly learn the representations and perform clustering in a unified framework [25, 31, 53, 55]. One successful deep approach to the subspace clustering problem is presented in [16], known as Deep Subspace Clustering (DSC), which employs a deep convolutional auto-encoder to learn latent representations and uses a novel self-expressive layer to enforce them to lie on a union of linear subspaces. The DSC model is further adopted by Deep Adversarial Subspace Clustering (DASC) method [55] to develop an adversarial architecture, consisting of a generator to produce subspaces and a discriminator to supervise the generator by evaluating the quality of the subspaces. More recently, [53] introduced an end-to-end trainable framework, named Self-Supervised Convolutional Subspace Clustering Network (S^{2}ConvSCN), which aims to

jointly learn feature representations, self-expression coefficients, and the clustering results to produce more accurate clusters.

Our approach can be seen as a generalization of the DSC algorithm [16] to the case that low-level and high-level information of the input data is utilized to produce more informative and discriminative subspace clustering representations. Moreover, we introduce a loss minimization problem that employs an initial clustering of the samples to effectively aggregate the knowledge gained from multi-level representations and to promote learning more accurate subspaces. Notice that although our work is close to DASC [55] and S^{2}ConvSCN [53] in the sense that it leverages a clustering of the samples to improve the feature learning procedure, we adopt a completely different strategy to incorporate the pseudo-label information into the problem.

It is noteworthy to emphasize that our approach may seem similar to the multi-view subspace clustering approaches [10, 23, 39, 52] as it aggregates information obtained from multiple modalities of the data to recover the clusters more precisely. However, it differs from them in the sense that our method leverages some connection layers to simultaneously learn multi-level deep representations and effectively fuse them to boost the clustering performance.

3. Problem Formulation

Let $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$ be a set of n sample points drawn from a union of K different subspaces in $\mathbb{R}^d$ that are not necessarily linear. An effective approach to cluster the samples is to transform them into a set of new representations that have linear relationships and satisfy the self-expressiveness property. Then, spectral clustering can be applied to recover the underlying clusters. To this end, the DSC algorithm [16] introduced a deep architecture consisting of a convolutional autoencoder with L layers to generate latent representations and a fully-connected linear layer inserted between the encoder and decoder to ensure the self-expressiveness property is preserved. Let $\mathcal{E}$ and $\mathcal{D}$, parameterized by Θ_e and Θ_d, denote the encoder and the decoder networks, respectively. Given that, the DSC algorithm proposed to solve the following optimization problem to learn desired representations and infer self-expression matrix $\mathbf{C}$

$$\underset{\Theta}{\text{minimize}} \ \|\mathbf{X}-\hat{\mathbf{X}}_\Theta\|_{\mathrm{F}}^2+\lambda\|\mathbf{Z}_{\Theta_e}-\mathbf{Z}_{\Theta_e}\mathbf{C}\|_{\mathrm{F}}^2+\gamma\|\mathbf{C}\|_p \quad (2a)$$

$$\text{subject to } \operatorname{diag}(\mathbf{C}) = \mathbf{0}, \quad (2b)$$

where $\lambda, \gamma > 0$ are fixed hyperparameters to control the importance of different terms and $\Theta = \{\Theta_e, \mathbf{C}, \Theta_d\}$ shows the network parameters. Matrix $\mathbf{Z}_{\Theta_e} \in \mathbb{R}^{\bar{d}\times n}$ indicates the latent representations where $\bar{d}$ is the dimension of the representations and $\mathbf{Z}_{\Theta_e} = \mathcal{E}(\mathbf{X};\Theta_e)$, and matrix $\hat{\mathbf{X}}_\Theta \in \mathbb{R}^{d\times n}$ denotes the reconstructed samples where $\hat{\mathbf{X}}_\Theta = \mathcal{D}(\mathcal{E}(\mathbf{X};\Theta_e)\mathbf{C};\Theta_d)$. The main goal of problem (2a) – (2b) is to compute the network parameters such that equality $\mathbf{Z}_{\Theta_e} = \mathbf{Z}_{\Theta_e}\mathbf{C}$ holds and the reconstructed matrix $\hat{\mathbf{X}}$ can well approximate the input data $\mathbf{X}$. [16] used the backpropagation technique followed by the spectral clustering algorithm to find the solution of the minimization problem (2a) – (2b) and determine the cluster memberships of the samples.

In what follows, we propose a new deep architecture that leverages information from different levels of the encoder to learn more informative representations and improve the subspace clustering performance.

4. Proposed Method

This section presents a detailed explanation of our proposed approach. As it can be seen from the problem (2a) – (2b), the DSC algorithm only relies on the latent variables $\mathbf{Z}_{\Theta_e}$ to perform clustering. Due to the fact that different layers of the encoder provide increasingly complex representations of the input data, it may be quite difficult to learn suitable subspace clustering representations from the output of the encoder. This provides a strong motivation to incorporate information from the lower layers of the encoder to boost the clustering performance. Towards this goal, our approach uses a new architecture which jointly benefits from the low-level and high-level information of the input data to learn more informative subspace clustering representations. The approach adds multiple fully-connected linear layers between the symmetrical layers of the encoder and the decoder to provide multiple paths of information flow through the network. These connection layers can not only enhance the ability of the network in extracting more complex information from the input data but also supervise the output of encoder layers to generate multiple sets of representations that satisfy the self-expressiveness property. Figure 2 illustrates an example architecture of our proposed approach. Observe that the representations learned at different levels of the encoder, denoted as $\{\mathbf{Z}_{\Theta_e}^l\}_{l=1}^L$, are input to the fully-connected linear layers and the outputs of these layers are fed into the decoder layers. This strategy allows the decoder to reuse the low-level information for producing more accurate reconstructions of the input data which in turn can improve the overall clustering performance.

We assume each fully-connected layer is associated with a self-expression matrix in the form of the summation of two matrices, where the first one is shared between the entire layers and the second one is a layer-specific matrix. The encoder, which can be seen as a mapping function from the input space to the representation space, aims to preserve the relations between the data samples at different levels of representations. Moreover, some samples may have stronger (or weaker) relations at different levels of the encoder. Define $\mathbf{C} \in \mathbb{R}^{n\times n}$ as the consistency matrix to capture the relational information shared between the encoder layers

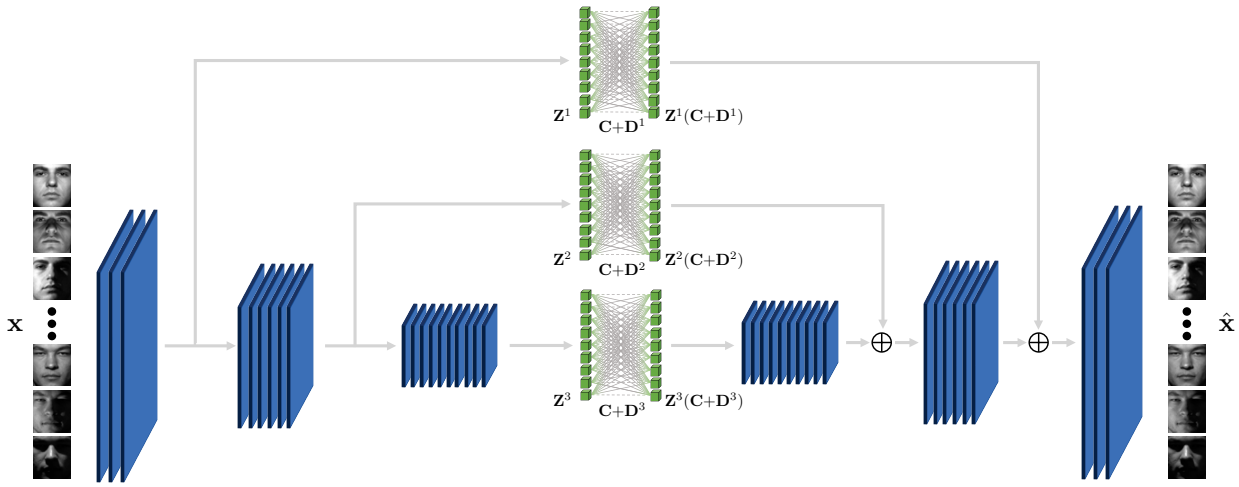


Figure 2: Architecture of the proposed multi-level representation learning model for $L = 3$. Observe that the representations learned at different levels of the encoder are fed into fully-connected linear layers to be used in the reconstruction procedure. Such strategy enables to combine low-level information from the early layers with high-level information from the deeper layers to produce more informative and robust subspace clustering representations. Each fully-connected layer is associated with a self-expression matrix formed from the summation of a coefficient matrix $\mathbf{C}$ shared between all layers and a distinctive matrix $\mathbf{D}^l$, $l \in \{1, \ldots, L\}$, which captures the unique information of each individual layer.

and $\{\mathbf{D}^l\}_{l=1}^L \in \mathbb{R}^{n\times n}$ as distinctive matrices to produce the unique information of the individual layers. Given that, we incorporate the following loss function to promote learning self-expressive representations

$$\mathcal{L}_{exp} = \sum_{l=1}^{L} \|\mathbf{Z}^l_{\Theta_e} - \mathbf{Z}^l_{\Theta_e}(\mathbf{C}+\mathbf{D}^l)\|_{\mathrm{F}}^2. \tag{3}$$

The above formulation is able to simultaneously model the shared information across different levels while considering the unique knowledge gained from each individual layer. This property allows to effectively leverage the information from the representations learned at multiple levels of the encoder and therefore is also particularly well-suited to the problem of multi-view subspace clustering [23].

The self-expression loss $\mathcal{L}_{exp}$ is employed to promote learning self-expressive feature representations at different levels of the encoder. To better accomplish this purpose, it is beneficial to adopt certain matrix norms for imposing desired structures on the elements of the distinctive matrices $\{\mathbf{D}^l\}_{l=1}^L$ and the consistency matrix $\mathbf{C}$. For the distinctive matrices, we use Frobenius norm to ensure the connectivity of the affinity graph associated with each fully-connected layer. For the consistency matrix $\mathbf{C}$, we employ ℓ_1-norm to generate sparse representations of the data. Ideally, it is desired to infer the consistency matrix and the distinctive matrices such that sample $\mathbf{x}_i$ is only expressed by a linear combination of the samples belonging to the same subspace as $\mathbf{x}_i$. To ensure these matrices obey the aforementioned desired structures, we propose to incorporate the following regularization terms

$$\mathcal{L}_{\mathbf{C}} = \|\mathbf{Q}^\top|\mathbf{C}|\|_1, \qquad \mathcal{L}_{\mathbf{D}} = \sum_{l=1}^{L} \|\mathbf{D}^l\|_{\mathrm{F}}^2, \tag{4}$$

where $\|.\|_1$ computes the sum of absolute values of its input matrix. Regularization term $\mathcal{L}_{\mathbf{C}}$ is used to incorporate the information gained from an initial pseudo-labels of the input data into the model. Let $\mathbf{Q} \in \mathbb{R}^{n\times K}$ be a membership matrix with its rows are one-hot vectors denoting the initial pseudo-labels assigned to the samples. The multiplication of $\mathbf{Q}^\top$ and $|\mathbf{C}|$ gives a matrix whose $(i,j)^{th}$ element shows the contribution of the samples assigned to the i^{th} subspace in reconstructing the j^{th} sample. Unlike the commonly used regularization $\|\mathbf{C}\|_1$ which imposes sparsity on the entire elements of the consistency matrix $\mathbf{C}$, $\mathcal{L}_{\mathbf{C}}$ promotes sparsity on the cluster memberships of the samples. In other words, it encourages each data to be reconstructed by the samples with the same pseudo-label and hence can smooth the membership predictions of the samples to different subspaces. Moreover, the regularization term $\mathcal{L}_{\mathbf{D}}$ promotes the elements of the distinctive matrices to be similar in value, which in turn can enhance the connectivity of the affinity graph associated with each fully-connected layer.

Combining the loss function (3) and the regularization terms $\mathcal{L}_{\mathbf{C}}$ and $\mathcal{L}_{\mathbf{D}}$ together with the reconstruction loss $\|\mathbf{X}-$

$\hat{\mathbf{X}}\|_{\mathrm{F}}^2$ leads to the following optimization problem that needs to be solved for training our proposed model

$$\underset{\Theta\cup\{\mathbf{D}^l\}_{l=1}^L}{\text{minimize}} \ \|\mathbf{X}-\hat{\mathbf{X}}_\Theta\|_{\mathrm{F}}^2+\lambda_1\sum_{l=1}^{L}\|\mathbf{Z}_{\Theta_e}^l-\mathbf{Z}_{\Theta_e}^l(\mathbf{C}+\mathbf{D}^l)\|_{\mathrm{F}}^2 + \lambda_2\|\mathbf{Q}^\top|\mathbf{C}|\|_1+\lambda_3\sum_{l=1}^{L}\|\mathbf{D}^l\|_{\mathrm{F}}^2 \tag{5a}$$

$$\text{subject to} \ \ \mathrm{diag}(\mathbf{C}+\mathbf{D}^l)=\mathbf{0}, \qquad l\in\{1,\ldots,L\}, \tag{5b}$$

where $\lambda_1,\lambda_2,\lambda_3>0$ are hyperparameters to balance the contribution of different losses. We adopt standard back-propagation technique to obtain the solution of problem (5a)–(5b). Once the solution matrices $\overset{*}{\mathbf{C}}$ and $\{\overset{*}{\mathbf{D}}^l\}_{l=1}^L$ are obtained, we can create a symmetric affinity matrix $\mathbf{W}\in\mathbb{S}_n$ of the following form

$$\mathbf{W}=\frac{\big|\overset{*}{\mathbf{C}}+\frac{1}{L}\sum_{l=1}^{L}\overset{*}{\mathbf{D}}^l\big|}{2}+\frac{\big|\overset{*}{\mathbf{C}}^\top+\frac{1}{L}\sum_{l=1}^{L}\overset{*}{\mathbf{D}}^{l^\top}\big|}{2}, \tag{6}$$

which shows the pairwise relations between the samples. Given that, the spectral clustering algorithm can be utilized to recover the underlying subspaces and cluster the samples to their respective subspaces.

Note that the pseudo-labels generated by spectral clustering can be leveraged to retrain the model and provide a more precise estimation of the subspaces. To this end, we assume the membership matrix $\mathbf{Q}$ is a variable and develop an iterative scheme to jointly learn the network parameters and matrix $\mathbf{Q}$. The approach starts from an initial $\mathbf{Q}$ (or equivalently an initial clustering of the input data) and alternatively runs the model for T epochs to train the network parameters $\Theta\cup\{\mathbf{D}^l\}_{l=1}^L$ and then updates the membership matrix. This training procedure is then repeated until the number of epochs reaches $\mathrm{maxIter}$. Different steps of our proposed scheme are delineated in detail in Algorithm 1.

Algorithm 1 Proposed Subspace Clustering Approach

Input: $\mathbf{X}$, $\mathbf{Q}$, T, $k=1$

repeat
  Update network parameters $\Theta\cup\{\mathbf{D}^l\}_{l=1}^L$ by solving (5a)–(5b) for one epoch
  if $k \bmod T=0$ **then**
    Form affinity matrix $\mathbf{W}$
    Apply spectral clustering to update $\mathbf{Q}$
  end if
  $k\leftarrow k+1$
until $k\leq \mathrm{maxIter}$

Output: $\mathbf{Q}$

Observe that Algorithm 1 can train the network parameters $\Theta\cup\{\mathbf{D}^l\}_{l=1}^L$ from scratch given the input matrices $\mathbf{X}$, $\mathbf{Q}$, and scalar T. However, several aspects of the algorithm such as convergence behavior and accuracy can be considerably improved by employing pre-trained models and using fine-tuning techniques to obtain initial values for the encoder and the decoder networks [16].

In the next section, we perform extensive experiments to corroborate the effectiveness of the proposed model. Also, we present a detailed explanation about the parameter settings, the pre-trained models, and the fine-tuning procedures used in our experiments.

5. Experiments

This section evaluates the clustering performance of our proposed method, termed MLRDSC, on four standard benchmark datasets for subspace clustering including two face image datasets (ORL and Extended Yale B) and two object image datasets (COIL20 and COIL100). Sample images from each of the datasets are illustrated in Figure 3. We perform multiple subspace clustering experiments on the datasets and compare the results against some baseline algorithms, including Low Rank Representation (LRR) [22], Low Rank Subspace Clustering (LRSC) [41], Sparse Subspace Clustering (SSC) [8], SSC with the pre-trained convolutional auto-encoder features (AE+SSC), Kernel Sparse Subspace Clustering (KSSC) [30], SSC by Orthogonal Matching Pursuit (SSC-OMP) [48], Efficient Dense Subspace Clustering (EDSC) [14], EDSC with the pre-trained convolutional auto-encoder features (AE+EDSC), Deep Subspace Clustering (DSC) [16], and Deep Adversarial Subspace Clustering (DASC) [55], Self-Supervised Convolutional Subspace Clustering Network (S^2ConvSCN) [53]. For the competitor methods, we directly collect the scores from the corresponding papers and some existing literature [16, 53].

Note that the subspace clustering problem is regarded as a specific clustering scenario which seeks to cluster a set of given unlabeled samples into a union of low-dimensional subspaces that best represent the sample data. In this sense, the subspace clustering approaches are basically different from the standard clustering methods that aim to group the samples around some cluster centers. Most of the subspace clustering literature revolves around using the spectral clustering technique to recover underlying subspaces from an affinity matrix, created over the entire samples. This can considerably increase the computational cost of these methods in comparison to the standard clustering approaches. As a consequence of this limitation, the benchmark datasets used for subspace clustering are generally smaller than that for the clustering task. In this work, we perform experiments on the aforementioned four datasets which are frequently used in the recent literature [8, 16, 53, 55] to evaluate the performance of the subspace clustering approaches.

In what follows, we first describe the training procedure

(a) Extended Yale B (b) ORL (c) COIL20 and COIL100

Figure 3: Example images of Extended Yale B, ORL, COIL20, and COIL100 datasets. The main challenges in the face image datasets, Extended Yale B and ORL, are illumination changes, pose variations and facial expression variations. The main challenges in the object image datasets, COIL20 and COIL100, are the variations in the view-point and scale.

used in our experiments. Then, we provide more details for each dataset separately and report the clustering performance of state-of-the-art methods.

5.1. Training Procedure

Following the literature [16, 55], for the convolutional layers, we use kernel filters with stride 2 in both dimensions and adopt rectified linear unit (ReLU) activation function. For the fully-connected layers, we use linear weights without considering bias or non-linear activation function. In order to train the model and obtain the affinity matrix, we follow the literature [16, 53, 55] and pass the entire samples into the model as a single batch. The Adam optimizer [18] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate 0.001 is used to train the network parameters. All experiments are implemented in PyTorch and the source code will be publicly available on the author's webpage.

As it is mentioned in [16], training the model from scratch is computationally expensive mainly because the samples are passed through the network as a single batch. To address this issue and following [16], we produce a pre-trained model by shortcutting all connection layers (i.e. $\mathbf{C} + \mathbf{D}^l = \mathbf{I}$ for $l \in \{1, \ldots, L\}$) and ignoring the self-expression loss term $\mathcal{L}_{exp}$. The resulting model is trained on the entire sample points and it can be utilized to initialize the encoder and the decoder parameters of our proposed architecture. We initialize the membership matrix $\mathbf{Q}$ to a zero matrix in all experiments ($\mathcal{L}_{\mathbf{C}} = 0$ for the first T epochs), although existing methods can be utilized to obtain better initialization. Moreover, we set each of the individual matrices $\mathbf{C}$ and $\{\mathbf{D}^l\}_{l=1}^{L}$ to a matrix with all elements equal to 0.0001. Notice that Algorithm 1 may fail to generate a convergent sequence of $\mathbf{Q}$ as it is terminated after $\mathrm{maxIter}$ epochs. One practical solution to handle this issue is to continue the training procedure until $\mathbf{Q}$ converges to a stable matrix [21].

5.2. Results

The results of all experiments are reported based on the clustering error which is defined to be the percentage of the misclustered samples to the entire sample points.

Extended Yale B: This dataset is used as a popular benchmark for the subspace clustering problem. It consists of 2432 frontal face images of size 192×168 captured from 38 different human subjects. Each subject has 64 images taken under different illumination conditions and poses. For computational purposes and following the literature [8, 16, 53], we downsample the entire images from their original size to 48×42.

We perform multiple experiments for a different number of human subjects $K \in \{10, 15, 20, 25, 30, 35, 38\}$ of the dataset to evaluate the sensitivity of MLRDSC with respect to increasing the number of clusters. By numbering the subjects from 1 to 38, we perform experiments on all possible K consecutive subjects and present the mean and median clustering errors of each $39 - K$ trials. Such experiments have been frequently performed in the literature [8, 16, 55, 53]. Through these experiments, we have employed an autoencoder model consisting of three stacked convolutional encoder layers with 10, 20, and 30 filters of sizes 5×5, 3×3, and 3×3, respectively. The parameters used in the experiments on this dataset are as follows: $\lambda_1 = 1 \times 10^{\frac{K}{10}-1}$, $\lambda_2 = 40$, $\lambda_3 = 10$, and we update the membership matrix $\mathbf{Q}$ in every $T = 100$ consecutive epochs. For the entire choices of K, we set the maximum number of epochs to 900. The clustering results on this dataset are reported in Table 1. Observe that MLRDSC achieves smaller errors than the competitor methods in all experiments, except for the mean of clustering error in case $K = 30$.

ORL: This dataset consists of 400 face images of size 112×92 from 40 different human subjects where each subject has

Table 1: Clustering error (%) of different methods on Extended Yale B dataset. The best results are in bold.

Measure	LRR	LRSC	SSC	AE+SSC	KSSC	SSC-OMP	EDSC	AE+EDSC	DSC	S^2ConvSCN	MLRDSC
10 subjects											
Mean	22.22	30.95	10.22	17.06	14.49	12.08	5.64	5.46	1.59	1.18	**1.10**
Median	23.49	29.38	11.09	17.75	15.78	8.28	5.47	6.09	1.25	1.09	**0.94**
15 subjects											
Mean	23.22	31.47	13.13	18.65	16.22	14.05	7.63	6.70	1.69	1.12	**0.91**
Median	23.49	31.64	13.40	17.76	17.34	14.69	6.41	5.52	1.72	1.14	**0.99**
20 subjects											
Mean	30.23	28.76	19.75	18.23	16.55	15.16	9.30	7.67	1.73	1.30	**0.99**
Median	29.30	28.91	21.17	16.80	17.34	15.23	10.31	6.56	1.80	1.25	**1.02**
25 subjects											
Mean	27.92	27.81	26.22	18.72	18.56	18.89	10.67	10.27	1.75	1.29	**1.13**
Median	28.13	26.81	26.66	17.88	18.03	18.53	10.84	10.22	1.81	1.28	**1.12**
30 subjects											
Mean	37.98	30.64	28.76	19.99	20.49	20.75	11.24	11.56	2.07	**1.67**	1.78
Median	36.82	30.31	28.59	20.00	20.94	20.52	11.09	10.36	2.19	1.72	**1.41**
35 subjects											
Mean	41.85	31.35	28.55	22.13	26.07	20.29	13.10	13.28	2.65	1.62	**1.44**
Median	41.81	31.74	29.04	21.74	25.92	20.18	13.10	13.21	2.64	1.60	**1.47**
38 subjects											
Mean	34.87	29.89	27.51	25.33	27.75	24.71	11.64	12.66	2.67	1.52	**1.36**
Median	34.87	29.89	27.51	25.33	27.75	24.71	11.64	12.66	2.67	1.52	**1.36**

10 images taken under diverse variation of poses, lighting conditions, and facial expressions. Following the literature, we downsample the images from their original size to 32×32. This dataset is challenging for subspace clustering due to the large variation in the appearance of facial expressions (shown in Figure 3) and since the number of images per each subject is quite small.

Through the experiment on ORL, we have adopted a network architecture consisting of three convolutional encoder layers with 3, 3, and 5 filters, all of size 3×3. Moreover, the parameter settings used in the experiment are as follows: $\lambda_1 = 5$, $\lambda_2 = 0.5$, $\lambda_3 = 1$, $T = 10$, and the maximum number of epochs is set to 420. The results of this experiment are presented in Table 2. It can be seen that MLRDSC outperforms all the competitor methods, except S^2ConvSCN which attains the smallest clustering error rate on ORL.

COIL20/COIL100: These two datasets are widely used for different types of clustering. COIL20 contains 1440 images captured from 20 various objects and COIL100 has 7200 images of 100 objects. Each object in either of the datasets has 72 images with black background taken at pose intervals of 5 degrees. The large viewpoint changes can pose serious challenges for the subspace clustering problem on these two dataset (Shown in in Figure 3).

For COIL20 and COIL100 datasets, the literature methods [16, 53, 55] mostly adopt one layer convolutional autoencoders to learn feature representations. This setting admits no connection layer and hence is not well-suited to our approach. To better demonstrate the advantages of MLRDSC, we use a two layers convolutional autoencoder model with 5 and 10 filters for performing experiment on COIL20 and adopt the same architecture with 20 and 30 filters for COIL100. The entire filters used in both experiments are of size 3×3. Moreover, the parameter settings for the datasets are as follows: 1) COIL20: $\lambda_1 = 20$, $\lambda_2 = 20$, $\lambda_3 = 5$, $T = 5$, and the maximum number of epochs is set to 50; 2) COIL100: $\lambda_1 = 20$, $\lambda_2 = 40$, $\lambda_3 = 10$, $T = 50$, and the maximum number of epochs is set to 350. The results on COIL20 and COIL100 datasets are shown in Table 2. Observe that our approach achieves better subspace clustering results on both datasets compared to the existing state-of-the-art methods.

According to the Tables 1 and 2, the deep subspace clustering methods, such as DSC, S^2ConvSCN, and MLRDSC, perform considerably well compared to the classical subspace clustering approaches on the benchmark datasets. This success can be attributed to the fact that deep models are able to efficiently capture the non-linear relationships between the samples and recover the underlying subspaces. Moreover, the results indicate that MLRDSC outperforms the DSC algorithm by a notable margin. This improvement can be resulted from the incorporation of a modified regularization term and the insertion of connection layers between the corresponding layers of the encoder and decoder. These layers enable the model to combine the information of different levels of the encoder to learn more favorable subspace clustering representations. It is noteworthy to mention that although our approach achieves better clustering results than the DSC method, it has more parameters to train, which in turn increases the computational burden of the model.

Ablation Study: To highlight the benefits brought by different components of our proposed model, we carry out an ablation study by evaluating a variant of our approach,

Table 2: Clustering error (%) of different methods on ORL, COIL20, and COIL100 datasets. The best results are in bold.

Dataset	LRR	LRSC	SSC	AE+SSC	KSSC	SSC-OMP	EDSC	AE+EDSC	DSC	DASC	S^{2}ConvSCN	MLRDSC
ORL	33.50	32.50	29.50	26.75	34.25	37.05	27.25	26.25	14.00	11.75	**10.50**	11.25
COIL20	30.21	31.25	14.83	22.08	24.65	29.86	14.86	14.79	5.42	3.61	2.14	**2.08**
COIL100	53.18	50.67	44.90	43.93	47.18	67.29	38.13	38.88	30.96	–	26.67	**23.28**

Table 3: Ablation study of our method in terms of clustering error (%) on Extended Yale B. The best results are in bold.

Measure	DSC-L2	DSC-L1	MLRDSC ($\|\mathbf{C}\|_1$)	MLRDSC
10 subjects				
Mean	1.59	2.23	**1.09**	1.10
Median	1.25	2.03	1.08	**0.94**
15 subjects				
Mean	1.69	2.17	0.98	**0.91**
Median	1.72	2.03	**0.99**	**0.99**
20 subjects				
Mean	1.73	2.17	**0.94**	0.99
Median	1.80	2.11	**0.94**	1.02
25 subjects				
Mean	1.75	2.53	**1.13**	**1.13**
Median	1.81	2.19	**1.12**	**1.12**
30 subjects				
Mean	2.07	2.63	1.84	**1.78**
Median	2.19	2.81	**1.35**	1.41
35 subjects				
Mean	2.65	3.09	1.49	**1.44**
Median	2.64	3.10	1.49	**1.47**
38 subjects				
Mean	2.67	3.33	1.40	**1.36**
Median	2.67	3.33	1.40	**1.36**

named MLRDSC ($\|\mathbf{C}\|_1$), which replaces $\mathcal{L}_{\mathbf{C}}$ with $\|\mathbf{C}\|_1$. In this sense, MLRDSC ($\|\mathbf{C}\|_1$) can be seen as a generalization of DSC-L1 (a variant of the DSC algorithm that utilizes regularization term $\|\mathbf{C}\|_1$ [16]) to a case that leverages multiple connection layers to learn multi-level subspace clustering representations. We perform experiments for different number of subjects K on Extended Yale B dataset and present the clustering results in Table 3. As the table indicates, inserting the connection layers between the symmetrical layers of the encoder and decoder can considerably improve the clustering performance of DSC-L1 algorithm. Moreover, comparing the results of MLRDSC and MLRDSC($\|\mathbf{C}\|_1$) confirms the positive effect of incorporating the regularization term $\mathcal{L}_{\mathbf{C}}$.

Sensitivity Analysis: we perform multiple experiments on the Extended Yale B dataset with various choices of hyperparameters (λ_1, λ_2, λ_3) to evaluate the sensitivity of the proposed approach to the choice of these parameters. The results of these experiments are reported in Table 4. Observe that the proposed approach exhibits a satisfactory performance for a wide range of these hyperparameters which demonstrates its generalization power.

Table 4: Sensitivity analysis of our method in terms of clustering error (%) on Extended Yale B. Triplet ($\bar{\lambda}_1$, $\bar{\lambda}_2$, $\bar{\lambda}_3$) corresponds to the parameter setting used to produce the results of Table 1.

λ_1	$\bar{\lambda}_1$	$\bar{\lambda}_1$	$\bar{\lambda}_1$	$\bar{\lambda}_1$	$\bar{\lambda}_1$	$0.1\bar{\lambda}_1$	$10\bar{\lambda}_1$
λ_2	$\bar{\lambda}_2$	$0.1\bar{\lambda}_2$	$100\bar{\lambda}_2$	$\bar{\lambda}_2$	$\bar{\lambda}_2$	$\bar{\lambda}_2$	$\bar{\lambda}_2$
λ_3	$\bar{\lambda}_3$	$\bar{\lambda}_3$	$\bar{\lambda}_3$	$0.1\bar{\lambda}_3$	$100\bar{\lambda}_3$	$\bar{\lambda}_3$	$\bar{\lambda}_3$
10 subjects							
Mean	1.10	1.11	1.15	1.09	1.06	2.52	1.08
Median	0.94	0.94	0.94	0.94	0.94	2.03	1.09
15 subjects							
Mean	0.91	0.92	0.99	0.92	0.94	1.28	0.97
Median	0.99	0.99	1.04	0.99	0.99	1.25	0.94
20 subjects							
Mean	0.99	0.99	1.00	0.99	1.00	0.83	0.94
Median	1.02	1.02	1.02	1.02	1.02	0.86	1.02
25 subjects							
Mean	1.13	1.13	1.14	1.13	1.14	1.51	1.13
Median	1.12	1.09	1.13	1.12	1.09	1.06	1.13
30 subjects							
Mean	1.78	2.26	2.42	1.86	1.97	2.43	1.70
Median	1.41	1.35	1.41	1.41	1.41	1.46	1.35
35 subjects							
Mean	1.44	1.45	1.45	1.45	1.45	1.45	1.50
Median	1.47	1.47	1.47	1.47	1.47	1.47	1.50
38 subjects							
Mean	1.36	1.36	1.32	1.36	1.36	1.32	1.40
Median	1.36	1.36	1.32	1.36	1.36	1.32	1.40

6. Conclusions

This paper presented a novel spectral clustering-based approach which uses a deep neural network architecture to address subspace clustering problem. The proposed method improves upon the existing deep approaches by leveraging information exploited from different levels of the networks to transform input samples into multi-level representations lying on a union of linear subspace. Moreover, it is able to use pseudo-labels generated by spectral clustering technique to effectively supervise the representation learning procedure and boost the final clustering performance. Experiments on benchmark datasets demonstrate that the proposed approach is able to efficiently handle clustering from the non-linear subspaces and it achieves better results compared to the state-of-the-art methods.

References

[1] M. Abavisani and V. M. Patel. Deep multimodal subspace clustering networks. *IEEE J. Sel. Topics Signal Process.*, 12(6):1601–1614, 2018.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. *Automatic subspace clustering of high dimensional data for data mining applications*, volume 27. ACM, 1998.

[3] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang. Diversity-induced multi-view subspace clustering. In *CVPR*, 2015.

[4] G. Chen and G. Lerman. Spectral curvature clustering (scc). *Int. J. Comput. Vis.*, 81(3):317–330, 2009.

[5] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2016.

[6] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan. Deep unsupervised clustering with Gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[7] E. L. Dyer, A. C. Sankaranarayanan, and R. G. Baraniuk. Greedy feature selection for subspace clustering. *J. Mach. Learn. Res.*, 14(1):2487–2517, 2013.

[8] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.

[9] P. Favaro, R. Vidal, and A. Ravichandran. A closed form solution to robust subspace estimation and clustering. In *CVPR*, 2011.

[10] H. Gao, F. Nie, X. Li, and H. Huang. Multi-view subspace clustering. In *ICCV*, 2015.

[11] K. Ghasedi Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *ICCV*, 2017.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[13] R. Heckel and H. Bölcskei. Robust subspace clustering via thresholding. *IEEE Trans. Inf. Theory*, 61(11):6320–6342, 2015.

[14] P. Ji, M. Salzmann, and H. Li. Efficient dense subspace clustering. In *WACV*, 2014.

[15] P. Ji, M. Salzmann, and H. Li. Shape interaction matrix revisited and robustified: Efficient subspace clustering with corrupted and incomplete data. In *ICCV*, 2015.

[16] P. Ji, T. Zhang, H. Li, M. Salzmann, and I. Reid. Deep subspace clustering networks. In *NeurIPS*, 2017.

[17] H. Kazemi, S. Soleymani, F. Taherkhani, S. Iranmanesh, and N. Nasrabadi. Unsupervised image-to-image translation using domain-specific variational information bound. In *NIPS*, 2018.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[19] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[20] F. Lauer and C. Schnörr. Spectral clustering of linear subspaces for motion segmentation. In *ICCV*, 2009.

[21] C.-G. Li and R. Vidal. Structured sparse subspace clustering: A unified optimization framework. In *CVPR*, pages 277–286, 2015.

[22] 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.*, 35(1):171–184, 2012.

[23] S. Luo, C. Zhang, W. Zhang, and X. Cao. Consistent and specific multi-view subspace clustering. In *AAAI*, 2018.

[24] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *ICANN*, 2011.

[25] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan. ClusterGAN: Latent space clustering in generative adversarial networks. *arXiv preprint arXiv:1809.03627*, 2018.

[26] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, 2002.

[27] E. Ntoutsi, K. Stefanidis, K. Rausch, and H.-P. Kriegel. Strength lies in differences: Diversifying friends for recommendations through subspace clustering. In *CIKM*, 2014.

[28] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004.

[29] V. M. Patel, H. Van Nguyen, and R. Vidal. Latent space sparse subspace clustering. In *ICCV*, 2013.

[30] V. M. Patel and R. Vidal. Kernel sparse subspace clustering. In *ICIP*, 2014.

[31] X. Peng, J. Feng, J. Lu, W.-Y. Yau, and Z. Yi. Cascade subspace clustering. In *AAAI*, 2017.

[32] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang. Structured autoencoders for subspace clustering. *IEEE Trans. Image Process.*, 27(10):5076–5086, 2018.

[33] X. Peng, S. Xiao, J. Feng, W.-Y. Yau, and Z. Yi. Deep subspace clustering with sparsity prior. In *IJCAI*, 2016.

[34] X. Peng, L. Zhang, and Z. Yi. Scalable sparse subspace clustering. In *CVPR*, 2013.

[35] P. Purkait, T.-J. Chin, A. Sadri, and D. Suter. Clustering with hypergraphs: the case for large hyperedges. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1697–1711, 2016.

[36] S. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(10):1832–1845, 2010.

[37] J. Shi and J. Malik. Normalized cuts and image segmentation. *Departmental Papers (CIS)*, page 107, 2000.

[38] M. Soltanolkotabi, E. J. Candes, et al. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[39] C. Tang, X. Zhu, X. Liu, M. Li, P. Wang, C. Zhang, and L. Wang. Learning joint affinity graph for multi-view subspace clustering. *IEEE Trans. Multimed.*, 2018.

[40] F. Tian, B. Gao, Q. Cui, E. Chen, and T.-Y. Liu. Learning deep representations for graph clustering. In *AAAI*, 2014.

[41] R. Vidal and P. Favaro. Low rank subspace clustering (LRSC). *Pattern Recognit. Lett.*, 43:47–61, 2014.

[42] S. Xiao, M. Tan, D. Xu, and Z. Y. Dong. Robust kernel low-rank representation. *IEEE Trans. Neural Netw. Learn. Syst.*, 27(11):2268–2281, 2015.

[43] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, 2016.

[44] X. Yang, C. Deng, F. Zheng, J. Yan, and W. Liu. Deep spectral clustering using dual autoencoder network. *arXiv preprint arXiv:1904.13113*, 2019.

[45] Y. Yang, J. Feng, N. Jojic, J. Yang, and T. S. Huang. ℓ^0-sparse subspace clustering. In *ECCV*, 2016.

[46] M. Yin, Y. Guo, J. Gao, Z. He, and S. Xie. Kernel sparse subspace clustering on symmetric positive definite manifolds. In *CVPR*, 2016.

[47] C. You, C.-G. Li, D. P. Robinson, and R. Vidal. Oracle based active set algorithm for scalable elastic net subspace clustering. In *CVPR*, 2016.

[48] C. You, D. Robinson, and R. Vidal. Scalable sparse subspace clustering by orthogonal matching pursuit. In *CVPR*, 2016.

[49] C. You and R. Vidal. Geometric conditions for subspace-sparse recovery. In *ICML*, 2015.

[50] A. Zhang, N. Fawaz, S. Ioannidis, and A. Montanari. Guess who rated this movie: Identifying users through subspace clustering. *arXiv preprint arXiv:1208.1544*, 2012.

[51] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao. Low-rank tensor constrained multiview subspace clustering. In *ICCV*, 2015.

[52] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao. Latent multi-view subspace clustering. In *CVPR*, 2017.

[53] J. Zhang, C.-G. Li, C. You, X. Qi, H. Zhang, J. Guo, and Z. Lin. Self-supervised convolutional subspace clustering network. In *CVPR*, 2019.

[54] T. Zhang, P. Ji, M. Harandi, W. Huang, and H. Li. Neural collaborative subspace clustering. *arXiv preprint arXiv:1904.10596*, 2019.

[55] P. Zhou, Y. Hou, and J. Feng. Deep adversarial subspace clustering. In *CVPR*, 2018.