

# Deep Closed-Form Subspace Clustering

Junghoon Seo  
Satrec Initiative, Republic of Korea  
sjh@satreci.com

Jamyoun Koo    Taegyun Jeon  
SI Analytics, Republic of Korea  
{jmkoo, tgjeon}@si-analytics.ai

## Abstract

We propose *Deep Closed-Form Subspace Clustering (DCFSC)*, a new embarrassingly simple model for subspace clustering with learning non-linear mapping. Compared with the previous deep subspace clustering (DSC) techniques, our DCFSC does not have any parameters at all for the self-expressive layer. Instead, DCFSC utilizes the implicit data-driven self-expressive layer derived from closed-form shallow auto-encoder. Moreover, DCFSC also has no complicated optimization scheme, unlike the other subspace clustering methods. With its extreme simplicity, DCFSC has significant memory-related benefits over the existing DSC method, especially on the large dataset. Several experiments showed that our DCFSC model had enough potential to be a new reference model for subspace clustering on large-scale high-dimensional dataset.

## 1. Introduction

In this paper, we tackle the problem of subspace clustering on high-dimensional and large-scale dataset. Subspace clustering [31] seeks to find clusters in the dataset by selecting the most relevant dimensions for each cluster separately. It has become an import topic in unsupervised learning and achieved great success in various computer vision tasks, such as face clustering [8], image segmentation [34], and motion segmentation [10, 27].

Recently, methods on subspace clustering based on sparse and low-rank representation [7, 35, 16, 33, 22, 36] have gotten attention. Many of these methods exploited self-expressiveness property [27, 6] of data drawn from a union of subspaces, i.e., the assumption that each data sample can be represented as a linear combination of other samples in the same subspace. The deep subspace clustering (DSC) network [12] is a deep auto-encoder based subspace clustering model to address the case of non-linear subspaces. The authors of DSC introduced the self-expressive layer to integrate self-expressiveness property into a deep neural network. This deep learning-based method was shown to outperform the other state-of-the-art subspace

clustering methods significantly. However, utilization of DSC is restricted to "shallow" models because the self-expressive layer requires a massive number of parameters.

In this paper, we propose a deep neural network to improve efficiency of self-expressiveness, which is termed **Deep Closed-Form Subspace Clustering (DCFSC)**. It consists of a closed form solution of the self-expressive layer motivated by  $EASE^R$  model [30], which showed that a similar Top- $N$  recommendation problem could be solved in closed form by a method of Lagrange multipliers. We modified the self-expressive layer from the parameterized fully-connected layer to a closed-form solution. In contrast to DSC, since the proposed self-expressive layer does not have any parameters for optimization, it is both memory-efficient and methodologically simple. To the best of our knowledge, this is the very first attempt that proposes to use closed-form solution to self-expressive layer. Furthermore, our model can use deeper neural networks for getting richer representation on subspace clustering.

We extensively evaluated our model on face clustering, using the Extended Yale B and ORL dataset for a small case, and on general object clustering, using COIL100 for a large case. Our experiments showed that DCFSC achieved comparable performance only using 0.25% ~ 0.44% parameters of DSC on a small case, and the state-of-the-art result on a large case.

## 2. Related Works

Subspace clustering problem usually is divided by two subproblems. The first subproblem is finding an affinity matrix from data. The second subproblem is clustering data points using the affinity matrix via normalized cuts [29] or spectral clustering [18]. Since there are already many good articles [21, 13, 31] that dealt comprehensively with classic subspace clustering methods, here we only deal with recent works on subspace clustering related with deep representation learning.

Several works [26, 25, 4] proposed a type of the methodology that representations learned by auto-encoder were forced to follow a specific conventional prior structure related with self-expression, e.g., Sparse Subspace Clustering

(SSC) [7] and Low-rank Representation (LRR) [16]. [15] proposed deep-encoder based row space recovery methodology to make conventional low-rank subspace clustering scalable and fast. [24] simultaneously learned a compact representation using a neural network, and [38] proposed combined methodology of a variant of K-subspace clustering [2] and deep auto-encoder to bypass the steps of constructing an affinity matrix and performing spectral clustering.

On the other hand, [12] firstly introduced deep subspace clustering network. The biggest contribution of [12] was that they firstly designed the self-expressive layer and corresponding loss function which models self-expressiveness property of data into deep auto-encoder. Since DSC showed great performance on various benchmarks, there have been many subsequent studies [41, 37, 40, 39] that tried to improve the DSC in several aspects. Deep adversarial subspace clustering [41] exploited GAN-like adversarial learning framework to supervise sample representation learning and subspace clustering. [37] proposed a dual self-supervision framework which exploited the output of spectral clustering to supervise the training of the feature learning module and the self-expression module. [40] introduced a new type of loss called distribution consistency loss to guide learning of distribution-preserving latent representation. [39] re-formulated subspace clustering as a classification problem, which in turn removed the spectral clustering step from the computations.

Despite the effectiveness and impact of DSC, the disadvantages of DSC have also been pointed out in some studies [38, 39]. The main disadvantage of DSC, which was commonly pointed out in these works, is that the memory footprint for training DSC is too high to access the subspace clustering problem for the large-scale dataset. This memory problem is caused by two factors of DSC. First, the self-expressive layer consists of  $N^2$  parameters for  $N$ -size dataset. Second, the process of clustering data points from the affinity matrix has a quiet high memory requirement. Neural Collaborative Subspace Clustering [39] tried to solve the latter, but did not face the former problem. Therefore, this paper is the primary work to solve the memory requirement problem of DSC’s self-expressive layer.

### 3. Proposed Framework

#### 3.1. Deep Subspace Clustering

Here, we firstly give a brief introduction on deep subspace clustering [12], which is one of key papers in this work. The core of DSC is joint training of deep auto-encoder and self-expressive layer. Let  $AE_{\Theta_{ae}} : \mathbb{R}^{N \times D} \rightarrow \mathbb{R}^{N \times d}$  denote the auto-encoder, which is parameterized

with  $\Theta_{ae}$ .<sup>1</sup>  $AE$  consists of two parts of feed-forward functions, a encoder  $Enc_{\Theta_{enc}} : \mathbb{R}^{N \times D} \rightarrow \mathbb{R}^{N \times d}$  and a decoder  $Dec_{\Theta_{dec}} : \mathbb{R}^{N \times d} \rightarrow \mathbb{R}^{N \times D}$ .  $Enc$  and  $Dec$  are parameterized with  $\Theta_{enc}$  and  $\Theta_{dec}$ , respectively. Let matrix  $X \in \mathbb{R}^{N \times D}$  represent a  $N$ -size whole dataset. Each row of  $X$  refers to each  $D$ -dimensional data point. Standard auto-encoder is trained to optimize (L2) reconstruction error  $L(X; \Theta_{ae})$ :

$$L(X; \Theta_{ae}) = \|X - AE(X)\|_F^2 = \|X - Dec(Enc(X))\|_F^2. \quad (1)$$

However, training with Equation 1, the latent representation  $Enc(X)$  is not guaranteed to have any beneficial property for subspace clustering.

For the guarantee, DSC utilizes self-expressiveness property of data drawn from union of linear subspaces [27, 6, 7]. Self-expressiveness property of set of points is that there exists a matrix  $C \in \mathbb{R}^{N \times N}$  which satisfies  $X = CX$  if each row data of  $X$  are drawn from one of the multiple linear subspaces.  $C$  is called self-representation coefficient matrix. If each subspace is independent with other subspace, self-representation coefficient matrix  $C$  has a block-diagonal structure [11]. With matrix norm constraint on  $C$ , finding optimal  $C$  under these two assumptions is formulated as the following:

$$\min_C \|C\|_p \quad \text{s.t.} \quad X = CX, \text{diag}(C) = 0. \quad (2)$$

Usually, complex high-dimensional data points in original data space itself do not satisfy self-expressiveness property so appropriate  $C$  cannot be found.

Instead of building assumption of self-expressiveness on data space, DSC enforces latent space of data  $Enc(X)$  to satisfy self-expressiveness property while training deep auto-encoder. Self-representation coefficient matrix is instantiated as trainable parameters of self-expressive layer  $SEL_{\Theta_{sel}} : \mathbb{R}^{N \times d} \rightarrow \mathbb{R}^{N \times d}$ .  $\Theta_{sel} \in \mathbb{R}^{N \times N}$  denotes parameters of self-expressive layer. Mapping by self-expressive layer is simply expressed as linear mapping among input data i.e.  $SEL(Y) = \Theta_{sel}Y$  where  $Y \in \mathbb{R}^{N \times d}$ . Under DSC framework,  $AE(X)$  is defined as  $Dec(SEL(Enc(X)))$ .  $\Theta_{ae} = \{\Theta_{enc}, \Theta_{dec}\}$  and  $\Theta_{sel}$  are jointly optimized with constraints and regularization derived from the self-expressiveness property (Equation 2):

$$\begin{aligned} L(X; \Theta_{ae}, \Theta_{sel}) = & \|X - Dec(SEL(Enc(X)))\|_F^2 \quad (3) \\ & + \lambda_1 \|\Theta_{sel}\|_p \\ & + \frac{\lambda_2}{2} \|Enc(X) - SEL(Enc(X))\|_F^2 \\ \text{s.t.} \quad & \text{diag}(\Theta_{sel}) = 0, \end{aligned}$$

<sup>1</sup>In the rest of the paper, when there is no confusion, the subscript that represent the learnable parameters is omitted sometimes for simplicity of notation. This applies not only to  $AE_{\Theta_{ae}}$ , but also to all parameterized functions.

---

**Algorithm 1** Deep Subspace Clustering [12]

**Input:** Data Matrix  $X$ ,  
Encoder with Pre-trained Parameters  $Enc_{\Theta_{enc}}$ ,  
Parameters of Self-expressive Layer  $\Theta_{sel}$ ,  
Decoder with Pre-trained Parameters  $Dec_{\Theta_{dec}}$ ,  
Hyper-parameters for Loss Weights  $\lambda_1, \lambda_2$ ,  
Number of Training Iteration  $EndStep$

**Output**  $\tilde{\Theta}_{sel}$

- 1:  $\Theta \leftarrow \{\Theta_{enc}, \Theta_{sel}, \Theta_{dec}\}$
- 2:  $n\_iter \leftarrow 0$
- 3: **while**  $n\_iter < EndStep$  **do**
- 4:    $X_{latent} \leftarrow Enc_{\Theta_{enc}}(X)$
- 5:    $X_{latent}' \leftarrow \Theta_{sel} X_{latent}$
- 6:    $X_{recon} \leftarrow Dec_{\Theta_{dec}}(X_{latent}')$
- 7:   **Update**  $\Theta$  to minimize Equation 3
- 8:    $n\_iter \leftarrow n\_iter + 1$
- 9:  $\tilde{\Theta}_{sel} \leftarrow \Theta_{sel}$
- 10: **return**  $\tilde{\Theta}_{sel}$

---

where  $p = 1$  or  $p = 2$  in [12]. In this work, we only consider  $p = 2$  case. This is because using L-2 norm makes optimization free from the diagonal constraint [11] and usually shows better performance than using L-1 norm. Algorithm 1 shows the whole training scheme of DSC.

After the network is trained, parameters of self-expressive layer  $\Theta_{sel}$  is used for constructing affinity matrix  $A \in \mathbb{R}^{N \times N}$ . This affinity matrix is then used for spectral clustering [18] to yield the final data clustering result. For building affinity matrix from the parameters of the self-expressive layer, the official implementation of DSC utilizes sparse subspace clustering (SSC) algorithm [7]. To cluster data points from the affinity matrix, spectral clustering method [18] is used.

### 3.2. Closed Form Solution of Self-Expressive Layer

In this section, we consider the following optimization problem:

$$\min_{\Theta_{sel}} \|X' - \Theta_{sel} X'\|_F^2 + \lambda \|\Theta_{sel}\|_F^2 \quad \text{s.t.} \quad \text{diag}(\Theta_{sel}) = 0. \quad (4)$$

Problem 4 is the partial problem of DSC's objective when  $X'$  is defined as  $Enc_{\Theta_{enc}}(X)$ . Specifically, Problem 4 is the problem excluding auto-encoding loss term (Equation 1) from the minimization problem of the DSC's objective function (Equation 3).

The optimization problem 4 is usually dealt in several works [19, 20, 5, 30] in field of Top- $N$  recommendation problem. Recently, [30] showed that this problem could be simply solved in closed form by method of Lagrange multipliers. Motivated from it, our main approach to optimize Equation 3 is to adopt this closed-form solution of Prob-

---

**Algorithm 2** Deep Closed-Form Subspace Clustering

**Input:** Data Matrix  $X$ ,  
Encoder with Parameters  $Enc_{\Theta_{enc}}$ ,  
Decoder with Parameters  $Dec_{\Theta_{dec}}$ ,  
Matrix Regularization Parameter  $\lambda$ ,  
Number of Training Iteration  $EndStep$

**Output** Self-representation Coefficient Matrix  $B$

- 1:  $\Theta \leftarrow \{\Theta_{enc}, \Theta_{dec}\}$
- 2:  $n\_iter \leftarrow 0$
- 3: **while**  $n\_iter < EndStep$  **do**
- 4:    $X_{latent} \leftarrow Enc_{\Theta_{enc}}(X)$
- 5:    $P \leftarrow \text{compute\_p}(X_{latent}, \lambda)$
- 6:    $B \leftarrow \text{compute\_b}(P)$
- 7:    $\bar{B} \leftarrow \text{Stop\_Gradients}(B)$
- 8:    $X_{latent}' \leftarrow \bar{B} X_{latent}$
- 9:    $X_{recon} \leftarrow Dec_{\Theta_{dec}}(X_{latent}')$
- 10:   **Update**  $\Theta$  to minimize Equation 1
- 11:    $n\_iter \leftarrow n\_iter + 1$

**return**  $B$

---

lem 4 and minimize only Equation 1, instead of minimizing Equation 3 by first-order methods.

Following derivation in Section 3.1 of [30], the closed-form solution of Equation 4 is given as the following:

$$B = I - P \cdot \text{diagMat}(\vec{1} \oslash \text{diag}(P)), \quad (5)$$

where  $P = (XX^T + \lambda I)^{-1}$ . In Equation 5,  $\text{diagMat}(\cdot)$ ,  $\vec{1} \oslash$ , and  $\text{diag}(\cdot)$  denote operation converting vector to diagonal matrix, a vector of ones, Hadamard division of matrices, and operation converting diagonal matrix to vector, in order. With reconfiguration of Equation 5, the solution can become more computationally efficient form:

$$B_{ij} = \begin{cases} 0 & \text{if } i = j \\ -\frac{P_{ij}}{P_{ii}} & \text{otherwise.} \end{cases} \quad (6)$$

### 3.3. Deep Closed-Form Subspace Clustering

Our DCFSC is a variant of DSC with closed form solution of self-expressive layer. Algorithm 2 describes how training procedure of DCFSC works. The main differences between Algorithm 1 and Algorithm 2 are indicated by magenta and blue, respectively. Two core steps of DCFSC,  $\text{compute\_p}(\cdot, \cdot)$  and  $\text{compute\_b}(\cdot)$ , are directly matched with Equation 5. Listing 1 is Tensorflow [1] implementation for  $\text{compute\_p}(\cdot, \cdot)$  and  $\text{compute\_b}(\cdot)$ . As the readers can see, DCFSC is easy to implement as much as DSC.

Compared with DSC (Algorithm 1), DCFSC does not retain  $N \times N$ -size parameters for self-representation coefficient matrix so does not need to optimize them. Moreover, there is no need to compute gradient over  $B$  or  $P$  because

```

1 import tensorflow as tf
2
3 def compute_p(encoded, coef_lambda):
4     # In: encoded (Tensor with shape [N, d])
5     # In: coef_lambda (float)
6     # Out: Tensor with shape [N, N]
7     encoded_t = tf.transpose(encoded)
8     gram_matrix = tf.matmul(encoded, encoded_t)
9     identity = tf.eye(encoded.shape[0])
10    gram_matrix += coef_lambda * identity
11    p = tf.linalg.inv(gram_matrix)
12    return p
13
14 def compute_b(p):
15     # In: p (Tensor with shape [N, N])
16     # Out: Tensor with shape [N, N]
17     diag_p = tf.linalg.diag_part(p)
18     b = p / (-diag_p[:, tf.newaxis])
19
20     zeros = tf.zeros(b.shape[0:-1])
21     b = tf.linalg.set_diag(b, zeros)
22    return b

```

**Listing 1** Implementation of *compute\_p* and *compute\_b* on Tensorflow.

closed form solution for self-representation coefficient matrix is directly derived from  $X_{latent}$  only via forward pass.

In case of small dataset such as ORL ( $N = 400$ ) and Extended Yale B ( $N = 2,432$ ), it results in little benefit over the existing DSC method. However, if size of dataset is relatively large like COIL-100 ( $N = 7,200$ ), our approach has a significant benefit on memory efficiency. On large datasets, advantages of DCFSC over DSC enable us to use deeper architecture to get better latent representations for subspace clustering. In contrast to DCFSC, DSC only allows shallow models to be used because of memory issue related with self-representation coefficient matrix.

## 4. Experiments

**Compared Methods and Performance Metric** For performance comparison among several subspace clustering methods, we adopted list of methods and benchmark results from the previous works [26, 12]: Low Rank Representation (LRR) [16], Low Rank Subspace Clustering (LRSC) [32], Sparse Subspace Clustering (SSC) [7], Kernel Sparse Subspace Clustering (KSSC) [23], SSC by Orthogonal Matching Pursuit (SSCOMP) [35], Efficient Dense Subspace Clustering (EDSC) [11], SSC with the pre-trained convolutional auto-encoder features (AE+SSC), EDSC with the pre-trained convolutional auto-encoder features (AE+EDSC), and Deep Subspace Clustering (DSC) [12]. Since the performances of DSC with L2 regularization were reported to be consistently better than those of L1 regularization, only performances of L2-regularized version of DSC were reported here. We also used the clustering er-

ror rate as metric for evaluating performance of each subspace clustering method, as same with [12]. We collected benchmark results of various methods from the DSC paper.

**Design of Experiments** We separated experiments into two cases by size of dataset: small  $N$  cases (Section 4.1) and large  $N$  case (Section 4.2).

The design of small  $N$  case experiments was to show performance of DCFSC under the very same settings of DSC paper. The only difference were the algorithm part of DSC and DCFSC. The other settings of experiments (e.g., model architecture, training procedure, and evaluation protocol) were same with ones of the original DSC paper. In terms of performance, it might be quiet unfavorable and unfair for DCFSC because DCFSC has much smaller model parameters than DSC in same architecture setting. Thus, design of these experiments was intended to answer how well DCFSC worked in exactly the same settings as DSC in its paper, regardless of superior point of DCFSC on memory efficiency. In the small  $N$  cases, Extended Yale B dataset [14] and ORL dataset [28] were used.

In contrast to small  $N$  cases, the experiment on large  $N$  case was designed to verify performance with full use of DCFSC’s memory efficiency. In the experiment, convolutional auto-encoder architecture, which was deeper than that used in the work of DSC, was used for implementation of our DCFSC. Note that this deeper architecture was quiet computationally intractable under the DSC method. Thus, this experiment was intended to show our DCFSC’s computational efficiency and the followed possibility of stronger representation learning. COIL-100 dataset [17] was used for large  $N$  case.

**System Environment** Implementation of DCFSC for experiments was done with minimum modification of public implementation of the DSC paper.<sup>2</sup> The Python version used was 3.5.2, and the Tensorflow version was 1.8.0. In addition, a single NVIDIA TESLA V100 GPU with 40 Intel Xeon E5-2698 CPUs were used for the experiment, and the CUDA and CuDNN version used were 9.0 and 7.1.4, respectively.

### 4.1. Small $N$ Case: E-YaleB and ORL

**Data Description** Both Extended Yale B (E-YaleB) [14] and ORL [28] are face databases. Tuples representing number of classes, number of images per class, total number of images, and size of images on E-YaleB and ORL dataset are (38, 64, 2432,  $192 \times 168$ ) and (40, 10, 400,  $112 \times 92$ ), respectively. The main difficulty of E-YaleB is known as extreme illumination, whereas the difficulty of ORL is known

<sup>2</sup><https://github.com/panji1990/Deep-subspace-clustering-networks>

| Layers   | Encoder-1    | Encoder-2    | Encoder-3    | Self-Expressive | Decoder-1    | Decoder-2    | Decoder-3    |
|--|--------------|--------------|--------------|-----------------|--------------|--------------|--------------|
| Deep Subspace Clustering (Total # of Parameters: 5,929,615)          |              |              |              |                 |              |              |              |
| Kernel Size  | $5 \times 5$ | $3 \times 3$ | $3 \times 3$ | -               | $3 \times 3$ | $3 \times 3$ | $5 \times 5$ |
| # of Channels  | 10           | 20           | 30           | -               | 30           | 20           | 10           |
| # of Parameters  | 260          | 1,820        | 5,430        | 5,914,624       | 5,420        | 1,810        | 251          |
| Deep Closed-Form Subspace Clustering (Total # of Parameters: 14,991) |              |              |              |                 |              |              |              |
| Kernel Size  | $5 \times 5$ | $3 \times 3$ | $3 \times 3$ | -               | $3 \times 3$ | $3 \times 3$ | $5 \times 5$ |
| # of Channels  | 10           | 20           | 30           | -               | 30           | 20           | 10           |
| # of Parameters  | 260          | 1,820        | 5,430        | 0               | 5,420        | 1,810        | 251          |

Table 1: Comparison on Network settings for Extended Yale B. DCFSC has only model parameters of  $\frac{14,991}{5,929,615} \times 100\% = 0.25\%$  as compared to DSC.

| Layers  | Encoder-1    | Encoder-2    | Encoder-3    | Self-Expressive | Decoder-1    | Decoder-2    | Decoder-3    |
|---|--------------|--------------|--------------|-----------------|--------------|--------------|--------------|
| Deep Subspace Clustering (Total # of Parameters: 160,702)         |              |              |              |                 |              |              |              |
| Kernel Size   | $5 \times 5$ | $3 \times 3$ | $3 \times 3$ | -               | $3 \times 3$ | $3 \times 3$ | $5 \times 5$ |
| # of Channels   | 5            | 3            | 3            | -               | 3            | 3            | 5            |
| # of Parameters   | 130          | 138          | 84           | 160,000         | 84           | 140          | 126          |
| Deep Closed-Form Subspace Clustering (Total # of Parameters: 702) |              |              |              |                 |              |              |              |
| Kernel Size   | $5 \times 5$ | $3 \times 3$ | $3 \times 3$ | -               | $3 \times 3$ | $3 \times 3$ | $5 \times 5$ |
| # of Channels   | 5            | 3            | 3            | -               | 3            | 3            | 5            |
| # of Parameters   | 130          | 138          | 84           | 0               | 84           | 140          | 126          |

Table 2: Comparison on Network settings for ORL. DCFSC has only model parameters of  $\frac{702}{160,702} \times 100\% = 0.44\%$  as compared to DSC.

as deformation and various pose. Like experiment setting of the DSC paper [12], images of E-YaleB were resized to  $48 \times 48$ , and images of ORL were resized to  $32 \times 32$ .

**Experiment Settings** For small  $N$  cases, we used the almost same neural architecture used in DSC with little modification. The only difference was that we removed the self-expressive layers from the DSC network and changed the training algorithm from Algorithm 1 to Algorithm 2. Table 1 and Table 2 show overall comparisons of number of parameters between DSC and DCFSC for two small  $N$  cases experiments. Note that we did not extensively search for any other optimal training hyper-parameter or neural architecture for DCFSC.

All other settings of experiments for E-YaleB dataset and ORL dataset were same with experiments in the DSC paper. To measure the robustness of the DCFSC model for various numbers of clusters, we measured performance on several  $K$  subjects in the E-YaleB dataset. Here, number of clusters  $K$  was  $\{10, 15, 20, 25, 30, 35, 38\}$  and each subject was set to have 64 face images. For ORL dataset, number of clusters was set to 40, just like the original subject number of the ORL dataset. For both E-YaleB and ORL, learning rate and matrix regularization parameter  $\lambda$  in the DCFSC are set as 0.001 and  $5e^5$ , respectively. The model weights in the DCFSC were initialized to the pre-trained weights

used in the DSC experiments. For fine-tuning stage, the DCFSC model was trained by  $50 + 25K$  epochs for each  $K$  in E-YaleB dataset and by 700 epochs in ORL dataset.

**Results and Discussions** In terms of the number of parameters, the DCFSC model had only 0.25% and 0.44%, compared with the DSC model. The module that occupied most of the parameters in the DSC was the self-expressiveness layer. DCFSC and DSC showed no significant difference in terms of memory requirements during training. For instance, the amount of GPU memory required in training of DSC on ORL dataset was 1,022MB, whereas in DCFSC it was 942MB. Therefore, in small  $N$  cases, it is hard to say that DCFSC has a great advantage in learning procedure over DSC.

Benchmark results of small  $N$  cases are shown in Table 3 and Figure 1a. For various number of clusters in the E-YaleB dataset, the DCFSC showed mean of clustering error rate of 6.13%. This was significantly worse than DSC’s mean error rate (2.67%), but it was much better than other hard baselines (11.64% or higher). On the ORL dataset, DCFSC resulted in an error rate of 14.8%, which was slightly worse than 14.0% of DSC (but not significantly). In short, in the small  $N$  cases, the DCFSC showed almost equal or worse performance than the DSC on nearly the same settings as the experiment in the DSC paper. However,

| Method      | LRR   | LRSC  | SSC   | AE+<br>SSC | KSSC  | SSC-<br>OMP | EDSC        | AE+<br>EDSC | DSC         | DCFSC       |
|-------------|-------|-------|-------|------------|-------|-------------|-------------|-------------|-------------|-------------|
| 10 subjects |       |       |       |            |       |             |             |             |             |             |
| Mean        | 22.22 | 30.95 | 10.22 | 17.06      | 14.49 | 12.08       | 5.64        | <u>5.46</u> | <b>1.59</b> | 5.72        |
| Median      | 23.49 | 29.38 | 11.09 | 17.75      | 15.78 | 8.28        | <u>5.47</u> | 6.09        | <b>1.25</b> | 5.63        |
| 15 subjects |       |       |       |            |       |             |             |             |             |             |
| Mean        | 23.22 | 31.47 | 13.13 | 18.65      | 16.22 | 14.05       | 7.63        | 6.70        | <b>1.69</b> | <u>5.33</u> |
| Median      | 23.49 | 31.64 | 13.40 | 17.76      | 17.34 | 14.69       | 6.41        | 5.52        | <b>1.72</b> | <u>5.10</u> |
| 20 subjects |       |       |       |            |       |             |             |             |             |             |
| Mean        | 30.23 | 28.76 | 19.75 | 18.23      | 16.55 | 15.16       | 9.30        | 7.67        | <b>1.73</b> | <u>4.93</u> |
| Median      | 29.30 | 28.91 | 21.17 | 16.80      | 17.34 | 15.23       | 10.31       | 6.56        | <b>1.80</b> | <u>4.92</u> |
| 25 subjects |       |       |       |            |       |             |             |             |             |             |
| Mean        | 27.92 | 27.81 | 26.22 | 18.72      | 18.56 | 18.89       | 10.67       | 10.27       | <b>1.75</b> | <u>4.92</u> |
| Median      | 28.13 | 26.81 | 26.66 | 17.88      | 18.03 | 18.53       | 10.84       | 10.22       | <b>1.81</b> | <u>5.00</u> |
| 30 subjects |       |       |       |            |       |             |             |             |             |             |
| Mean        | 37.98 | 30.64 | 28.76 | 19.99      | 20.49 | 20.75       | 11.24       | 11.56       | <b>2.07</b> | <u>5.35</u> |
| Median      | 36.82 | 30.31 | 28.59 | 20.00      | 20.94 | 20.52       | 11.09       | 10.36       | <b>2.19</b> | <u>5.52</u> |
| 35 subjects |       |       |       |            |       |             |             |             |             |             |
| Mean        | 41.85 | 31.35 | 28.55 | 22.13      | 26.07 | 20.29       | 13.10       | 13.28       | <b>2.65</b> | <u>5.93</u> |
| Median      | 41.81 | 31.74 | 29.04 | 21.74      | 25.92 | 20.18       | 13.10       | 13.21       | <b>2.64</b> | <u>5.96</u> |
| 38 subjects |       |       |       |            |       |             |             |             |             |             |
|             | 34.87 | 29.89 | 27.51 | 25.33      | 27.75 | 24.71       | 11.64       | 12.66       | <b>2.67</b> | <u>6.13</u> |

Table 3: Clustering error (in %) on Extended Yale B. The lower the better. Lower is better. The **bold** and underlined text refer to the 1st and 2nd ranked score, respectively.

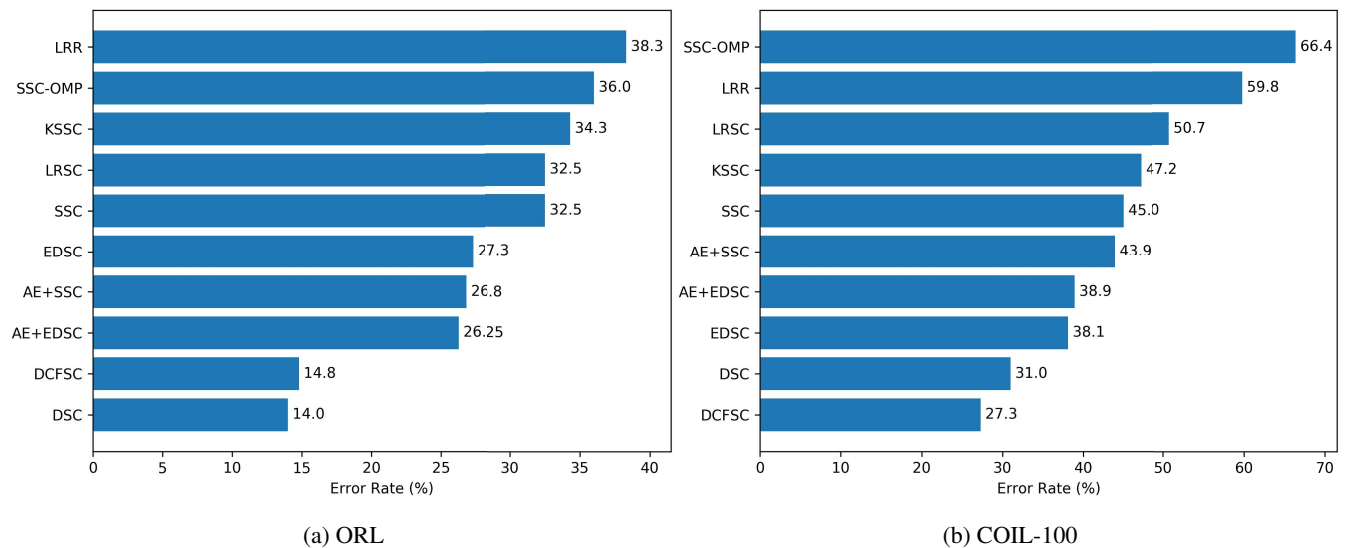


Figure 1: Subspace clustering error on ORL and COIL100. Methods are sorted in descending order of error. Lower is better. For both benchmarks of DCFSC, mean of ten trails is reported.

| Deep Subspace Clustering (Total # of Parameters: 51,842,600) |              |  |  |                 |  |  |              |  |  |  |  |
|--|--------------|--|--|-----------------|--|--|--------------|--|--|--|--|
| Layers   | Encoder-1    |  |  | Self-Expressive |  |  | Decoder-1    |  |  |  |  |
| Kernel Size  | $5 \times 5$ |  |  | -               |  |  | $5 \times 5$ |  |  |  |  |
| Stride Size  | 2            |  |  | -               |  |  | 2            |  |  |  |  |
| # of Channels  | 50           |  |  | -               |  |  | 50           |  |  |  |  |
| # of Parameters  | 1,300        |  |  | 51,840,000      |  |  | 1,300        |  |  |  |  |

| Deep Closed-Form Subspace Clustering (Total # of Parameters: 81,913) |              |              |              |              |              |                 |              |              |              |              |              |
|--|--------------|--------------|--------------|--------------|--------------|-----------------|--------------|--------------|--------------|--------------|--------------|
| Layers   | Enc-1        | Enc-2        | Enc-3        | Enc-4        | Enc-5        | Self-Expressive | Dec-1        | Dec-2        | Dec-3        | Dec-4        | Dec-5        |
| Kernel Size  | $5 \times 5$ | $3 \times 3$ | $3 \times 3$ | $3 \times 3$ | $1 \times 1$ | -               | $1 \times 1$ | $3 \times 3$ | $3 \times 3$ | $3 \times 3$ | $5 \times 5$ |
| Stride Size  | 1            | 2            | 1            | 2            | 1            | -               | 1            | 2            | 1            | 2            | 1            |
| # of Channels  | 24           | 24           | 48           | 48           | 72           | -               | 72           | 48           | 48           | 24           | 24           |
| # of Parameters  | 696          | 5,280        | 10,560       | 20,928       | 3,528        | 0               | 3,648        | 20,928       | 10,464       | 5,280        | 601          |

Table 4: Comparison on Network settings for COIL-100. DCFSC has only model parameters of  $\frac{81,913}{51,842,600} \times 100\% = 0.16\%$  as compared to DSC. Note that on the architecture used in DCFSC batch normalization layers [9] were used except for the last layer of the encoder and decoder.

the performance of DCFSC was dominant over all other existing baselines except DSC. The reason that DCFSC was inferior in performance to DSC might be that the neural architecture was not deep enough to yield the potential self-expression directly from data. Therefore, there still is room for improvement in DCFSC performance, like searching for an optimal architecture or hyper-parameter. These results, however, still show that convergence is experimentally guaranteed even in small  $N$  cases.

#### 4.2. Large $N$ Case: COIL-100

**Data Description** For large  $N$  case, COIL-100 dataset [17], which is a object database, was used to measure the performance of object clustering. On COIL-100 dataset, number of classes, number of images per class, total number of images, and size of images are 100, 72, 7200, and  $128 \times 128$ , respectively. The main difficulties of dealing with COIL-100 dataset are known as deformation and rotation. For consistency with previous studies [3, 12], images of COIL-100 were resized to  $32 \times 32$ .

**Experiment Settings** The model architecture, used in the original DSC work [12] for COIL-100 dataset, was a very shallow auto-encoder structure consisting of one encoder layer, a self-expressive layer, and one decoder layer. This was because the number of parameters that the self-expressive layer should retrain in case of large  $N$  was too huge to adopt deeper neural architecture due to memory problems. We used much deeper auto-encoder architecture to show that DCFSC could have tremendous advantages in this situation. Table 4 shows the difference between model architectures of DSC and DCFSC, used in the experiments. Unlike DSC, five encoder layers and five decoder layers

were used in DCFSC.

The number of clusters was set to 100, which was equal to the number of subjects in COIL-100. Learning rate and matrix regularization parameters were set to 0.001 and 10, respectively. While DSC had pre-trained model weights, the architecture of DCFSC in this experiment did not have such pre-trained model weights because new architecture was designed for training the DCFSC model. Thus, in the large  $N$  case experiment DCFSC was trained from scratch without pre-training. Because the DCFSC model did not use pre-training weights, it was trained for 175 epochs, which was longer than 120 epochs used in the DSC. All other experiment settings were same as in DSC’s ones.

**Results and Discussions** Most of the model parameters of DSC, which were used for large  $N$  case experiment, belonged to the self-expressive layer, and this tendency was much greater than in small  $N$  cases. This was because size of the self-expressive layer is proportional to the square of dataset size. In reality, the DSC model for COIL-100 required huge amount of self-expressive layer parameters ( $7,200 \times 7,200 = 51,840,000$ ). By storing these parameters in double precision, simply maintaining these parameters required about 3.2 GB of memory space. Furthermore, about 8.6GB of GPU memory was required to train the very shallow COIL-100 model presented in the DSC paper. This drawback of DSC made it be not able to use deeper architecture. Therefore, performance constraint of DSC on large dataset was practically inevitable. Our DCFSC is free from these limitations of DSC model. Number of parameters of the auto-encoder used in DCFSC was about 32 times of the DSC auto-encoder. Despite using much deeper architecture than the one used in the DSC, the GPU memory require-

ment to train it was 10.8GB, which was only 26% higher than the DSC architecture. This means that the DCFSC can actually use a deeper architecture than the DSC.

Figure 1b shows several benchmark results on COIL-100 dataset. Our DCFSC model showed a clustering error of 27.3% and outperformed all other models including DSC, without pre-training. These results reveal the performance advantage of DCFSC on larger dataset. In addition, they implies that deep learning based subspace clustering still has room to benefit from learning richer latent representations through deeper architecture. It is also noteworthy that performance of the DCFSC in this experiment was reported to be comparable with 26.6%, which was reported in the study [37] combining more sophisticated methodologies such as self-supervised learning with the DSC. Since DCFSC is easy to combine with the more advanced models [41, 37, 40, 39] of DSC, there is possibility of further enhancing the performance of these modified models with deeper neural architecture.<sup>3</sup>

### 4.3. Effect of $\lambda$ : Case of COIL-100

We further investigated the effect of selection of matrix regularization parameter  $\lambda$  on performance of DCFSC. This is to see how sensitive DCFSC performance is to the choice of  $\lambda$ , or how robust it is. For benchmarking, performances were reported by changing  $\lambda$  from 1 to  $1e^6$  in multiples of 10 in the same settings as the COIL-100 dataset (Section 4.2).

Figure 2 shows variation of subspace clustering error in COIL-100 with different selection of  $\lambda$ . It can be seen that choosing  $\lambda$  from 1 to 100 guaranteed better performance than DSC, and selection from 1 to 10 gave the best performance. Conversely, too large  $\lambda$  ( $\geq 1e^3$ ) degenerated performance. On the other hand, with fine-tuning the pre-trained model in small  $N$  cases (Section 4.1), a relatively large  $\lambda$  ( $5e^5$ ) was the appropriate choice for stable convergence. Therefore, the effect of  $\lambda$  selection on performance and the following optimal  $\lambda$  selection method need to be investigated further in terms of size of dataset, presence or absence of pre-training, and so on.

## 5. Conclusions

In this paper, we firstly propose a variant of the existing DSC method, which does not require retaining parameters of self-expressive layer. We call our method **Deep Closed-Form Subspace Clustering (DCFSC)** because it is motivated by recently proposed closed form of shallow auto-encoder model. Our DCFSC has advantages in methodological simplicity and memory efficiency compared to DSC.

<sup>3</sup>It is also remarkable that in 'Deep Adversarial Subspace Clustering' paper [41] the proposed model could not be used to try experiment in COIL-100, even with a very shallow auto-encoder model. It was also due to a memory shortage problem.

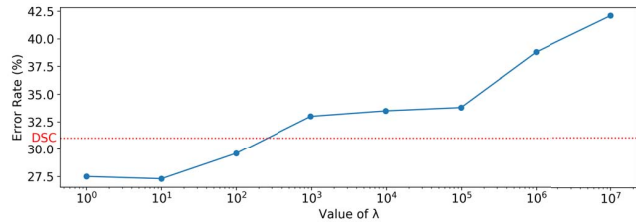


Figure 2: Effects of  $\lambda$  on clustering performance of COIL-100 experiments. Performance of DSC baseline is shown as red. For all benchmarks, mean of ten trials is reported.

Experiments on several benchmarks give two conclusions with regard to DCFSC. First, the DCFSC model could be trained and converged despite the disadvantage of much less model parameters even in small datasets under the same settings as DSC. Second, in large dataset, DCFSC could take advantage of memory and eliminate the performance limitations of DSC. Considering these strengths, we believe that DCFSC can be regarded as a model remedying the shortcomings of the existing DSC.

## References

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *USENIX Symposium on Operating Systems Design and Implementation*, pages 265–283, 2016. 3
- [2] Laura Balzano, Arthur Szlam, Benjamin Recht, and Robert Nowak. K-subspaces with missing data. In *IEEE Statistical Signal Processing Workshop*, pages 612–615. IEEE, 2012. 2
- [3] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2010. 7
- [4] Yuanyuan Chen, Lei Zhang, and Zhang Yi. Subspace clustering using a low-rank constrained autoencoder. *Information Sciences*, 424:27–38, 2018. 1
- [5] Yao Cheng, Liang Yin, and Yong Yu. Lorslim: Low rank sparse linear methods for top-n recommendations. In *IEEE International Conference on Data Mining*, pages 90–99. IEEE, 2014. 3
- [6] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2790–2797. IEEE, 2009. 1, 2
- [7] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013. 1, 2, 3, 4
- [8] Jeffrey Ho, Ming-Hsuan Yang, Jongwoo Lim, Kuang-Chih Lee, and David Kriegman. Clustering appearances of objects under varying illumination conditions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–18, 2003. 1



- [9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015. 7
- [10] Pan Ji, Hongdong Li, Mathieu Salzmann, and Yiran Zhong. Robust multi-body feature tracker: a segmentation-free approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3843–3851, 2016. 1
- [11] Pan Ji, Mathieu Salzmann, and Hongdong Li. Efficient dense subspace clustering. In *IEEE Winter Conference on Applications of Computer Vision*, pages 461–468. IEEE, 2014. 2, 3, 4
- [12] Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. In *Advances in Neural Information Processing Systems*, pages 24–33, 2017. 1, 2, 3, 4, 5, 7
- [13] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, 3(1):1, 2009. 1
- [14] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 684–698, 2005. 4
- [15] Jun Li and Hongfu Liu. Projective low-rank subspace clustering via learning deep encoder. In *International Joint Conference on Artificial Intelligence*, 2017. 2
- [16] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2012. 1, 2, 4
- [17] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-100), 1996. 4, 7
- [18] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2002. 1, 3
- [19] Xia Ning and George Karypis. Slim: Sparse linear methods for top-n recommender systems. In *IEEE International Conference on Data Mining*, pages 497–506. IEEE, 2011. 3
- [20] Xia Ning and George Karypis. Sparse linear methods with side information for top-n recommendations. In *ACM Conference on Recommender systems*, pages 155–162. ACM, 2012. 3
- [21] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004. 1
- [22] Vishal M Patel, Hien Van Nguyen, and René Vidal. Latent space sparse and low-rank subspace clustering. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):691–701, 2015. 1
- [23] Vishal M Patel and René Vidal. Kernel sparse subspace clustering. In *IEEE International Conference on Image Processing*, pages 2849–2853. IEEE, 2014. 4
- [24] Xi Peng, Jiashi Feng, Jiwen Lu, Wei-Yun Yau, and Zhang Yi. Cascade subspace clustering. In *AAAI Conference on Artificial Intelligence*, 2017. 2
- [25] Xi Peng, Jiashi Feng, Shijie Xiao, Wei-Yun Yau, Joey Tianyi Zhou, and Songfan Yang. Structured autoencoders for subspace clustering. *IEEE Transactions on Image Processing*, 27(10):5076–5086, 2018. 1
- [26] Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. Deep subspace clustering with sparsity prior. In *International Joint Conference on Artificial Intelligence*, 2016. 1, 4
- [27] Shankar R Rao, Roberto Tron, René Vidal, and Yi Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 1, 2
- [28] Ferdinando S Samaria and Andy C Harter. Parameterisation of a stochastic model for human face identification. In *IEEE Workshop on Applications of Computer Vision*, pages 138–142. IEEE, 1994. 4
- [29] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000. 1
- [30] Harald Steck. Embarrassingly shallow autoencoders for sparse data. In *ACM International World Wide Web Conference*, pages 3251–3257. ACM, 2019. 1, 3
- [31] René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011. 1
- [32] René Vidal and Paolo Favaro. Low rank subspace clustering (lrs). *Pattern Recognition Letters*, 43:47–61, 2014. 4
- [33] Yu-Xiang Wang, Huan Xu, and Chenlei Leng. Provable subspace clustering: When lrr meets ssc. In *Advances in Neural Information Processing Systems*, pages 64–72, 2013. 1
- [34] Allen Y Yang, John Wright, Yi Ma, and S Shankar Sastry. Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, 110(2):212–225, 2008. 1
- [35] Chong You, Daniel Robinson, and René Vidal. Scalable sparse subspace clustering by orthogonal matching pursuit. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3918–3927, 2016. 1, 4
- [36] Changqing Zhang, Huazhu Fu, Si Liu, Guangcan Liu, and Xiaochun Cao. Low-rank tensor constrained multiview subspace clustering. In *IEEE International Conference on Computer Vision*, pages 1582–1590, 2015. 1
- [37] Junjian Zhang, Chun-Guang Li, Chong You, Xianbiao Qi, Honggang Zhang, Jun Guo, and Zhouchen Lin. Self-supervised convolutional subspace clustering network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 8
- [38] Tong Zhang, Pan Ji, Mehrtash Harandi, Richard Hartley, and Ian Reid. Scalable deep k-subspace clustering. In *Asian Conference on Computer Vision*, pages 466–481. Springer, 2018. 2
- [39] Tong Zhang, Pan Ji, Mehrtash Harandi, Wen-bing Huang, and Hongdong Li. Neural collaborative subspace clustering. In *International Conference on Machine Learning*, 2019. 2, 8
- [40] Lei Zhou, Xiao Bai, Dong Wang, Xianglong Liu, Jun Zhou, and Edwin Hancock. Latent distribution preserving deep

subspace clustering. In *International Joint Conference on Artificial Intelligence*, 2019. [2](#), [8](#)

- [41] Pan Zhou, Yunqing Hou, and Jiashi Feng. Deep adversarial subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1596–1604, 2018. [2](#), [8](#)