

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,300

Open access books available

170,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



Chapter

# Scaling Subspace-Driven Approaches Using Information Fusion

*Sally Ghanem and Hamid Krim*

## Abstract

In this work, we seek to exploit the deep structure of multi-modal data to robustly exploit the group subspace distribution of the information using the Convolutional Neural Networks (CNNs) formalism. Upon unfolding the set of subspaces constituting each data modality, and learning their corresponding encoders, an optimized integration of the generated inherent information is carried out to yield a characterization of various classes. Referred to as deep Multimodal Robust Group Subspace Clustering (DRoGSuRe), this approach is compared against the independently developed state-of-the-art approach named Deep Multimodal Subspace Clustering (DMSC). Experiments on different multimodal datasets show that our approach is competitive and more robust in the presence of noise.

**Keywords:** sparse learning, computer vision, unsupervised classification, subspace clustering, multi-modal sensor data

## 1. Introduction

Unsupervised learning is a very challenging topic in Machine Learning (ML) and involves the discovery of hidden patterns in data for inference with no prior given labels. Reliable clustering techniques will save time and effort required for classifying/labeling large datasets that might have thousands of observations. Multi-modal data, increasingly in need for complex application problems, have become more accessible with recent advances in sensor technology, and of pervasive use in practice. The plurality of sensing modalities in our applications of interest, provides diverse and complementary information, necessary to capture the salient characteristics of data and secure their unique signature. A principled combination of the information contained in the different sensors and at different scales is henceforth pursued to enhance understanding of the distinct structure of the various classes of data. The objective of this work is to develop a principled multi-modal framework for object clustering in an unsupervised learning scenario. We extract key class-distinct features-signatures from each data modality using a CNNs encoder, and we subsequently non-linearly combine those features to generate a discriminative

characteristic feature. In so doing, we work on the hypothesis that each data modality is approximated by a Union of low dimensional Subspaces which highlights underlying hidden features. The UoS structure is unveiled by pursuing sparse self-representation of the given data modality. The subsequent aggregation of the multi-modal subspace structures yields a jointly unified characteristic subspace for each class.

### **1.1 Related work**

Subspace clustering has been introduced as an efficient way for unfolding union of low-dimensional subspaces underlying high dimensional data. Subspace clustering has been extensively studied in computer vision due to the vast availability of visual data as in [1–4]. This paradigm has broadly been adopted in many applications such as image segmentation [5], image compression [6], and object clustering [7].

Uncovering the principles and laying out the fundamentals for multi-modal data has become an important topic in research in light of many applications in diverse fields including image fusion [8], target recognition [9–12], speaker recognition [13], and handwriting analysis [14]. Convolutional neural networks have been widely used on multi-modal data as in [15–19]. A multi-modal subspace clustering-inspired approach was also proposed in [20]. The emphasis of our formulation results in a different optimization problem, as the multi-modal sensing seeks to not only account for the private information which provides the complementarity of the sensors, but also the common and hidden information. This yields, as an end result, a different network structure than that of [20] with a different application space inspiration. In addition, the robustness of fusing multi-modal sensor data each with its distinct intrinsic structure, is addressed along with a potential scaling for viability. A thorough comparison of our results to the multimodal fusion network in [21] is carried out, with a demonstration of resilient fusion under a variety of limiting scenarios including limited sensing modalities (sensor failures). In [22], the authors proposed a deep multi-view subspace clustering approach that combined global and local structures to help achieve a small distance between samples of the same cluster and make samples in different clusters of different views farther. To that end, they used a discriminative constraint between different views. The discriminative constraint is based on the Hadamard product between the features extracted by the convolutional auto-encoder for the different views. In contrast, our approach is based on the minimizing the group norm, which we proved with a derivation in earlier work [23] and entails a smaller angle between the different subspaces across all modalities, thus promoting the goal of obtaining a common latent space. Moreover, minimizing the group norm also provides as well as group sparse solution along data modalities. Sun et al. [24] proposed a deep trainable multi-view subspace clustering method, named self-supervised deep multi-view subspace clustering (S2DMVSC) that learns the common latent subspace using two losses: spectral clustering loss and classification loss in order to denoise the imperfect correlations among data points.

In this paper, we prove that our formulation, which is based on the group norm of the self-representation matrices and the commutation loss between them, provides a natural way to fuse multi-modal data by employing the self-representation matrix as an embedding for each data modality, making our approach robust under different types of potential limitations. It is good to note that our proposed approach secures the individual sensor data-points relations resulting in more flexibility for each data modality.

## 1.2 Contributions

Building on the work of Deep Subspace Clustering (DSC) [25], we propose a new and principled multi-modal fusion approach which accounts for a sensor's capacity to house private and unique information about some observed data as well as that information which is likely also captured and hence common to other sensors. This is accounted for in our robust fusion formulation for multi-modal sensor data. Unveiling the complex UoS of multi-modal data also requires us to account for scaling in our proposed formulation and solution, which in turn invokes the learning of multiple/deep scale Convolutional Neural Networks. Our proposed Multi-modal fusion approach, by virtue of each sensor information structure (i.e., private plus shared) seeks to enhance and robustify the subspace approximation of shared information for each of the sensors, thus yielding a parallel bank of UoS for each of the sensors. The robust Deep structure effectively achieves scaling while securing structured representation for unsupervised inference. We compare our approach to a well-known deep multimodal network [21] which was also based on [25].

In our proposed approach, we thus define the latent space in a way that safeguards the individual sensor private information which hence dedicates more degrees of freedom to each of the sensors. In contrast to the approach in [21]. In our evaluation, we use two recently released data sets each of which we partition into learning and validation subsets. The learned UoS structure for each of the data sets is then utilized to classify new observed data points, which illustrates the generalization power of the proposed approach. Different scenarios with corresponding additive noise to either the training set or the testing set, or both, were used to thoroughly investigate the robustness, and resilience of the clustering approach performance. Experimental results confirm a significant improvement for our Deep Robust Group Subspace Recovery network (DRoGSuRe) under numerous limiting scenarios and demonstrate robustness under these conditions.

The balance of the paper is organized as follows, in Section 2, we provide the problem formulation, background along with the derivation for our proposed approach, Deep Robust Group Subspace Recovery (DRoGSuRe). In Section 3, we describe the attributes of the proposed approach and contrast it to Deep Multimodal Subspace Clustering algorithm (DMSC). In Sections 4 and 5, we present a substantive validation along with experimental results of our approach, while Section 6 provides concluding remarks.

## 2. Deep robust group subspace clustering

### 2.1 Problem formulation

We assume having a set of data observations, each represented as a  $m$  dimensional vector  $\mathbf{x}_k(\mathbf{t}) \in \mathbb{R}^m$ , where  $k = 1, 2, \dots, n$ . Moreover, we consider having  $T$  data modalities, indexed by  $\mathbf{t} = 1, 2, 3, \dots, T$ . Each data modality can then be described as  $\mathbf{x}_k(\mathbf{t}) \in \mathbb{R}^m$ , where  $\mathbf{X}(\mathbf{t}) = [\mathbf{x}_1(\mathbf{t}) \ \mathbf{x}_2(\mathbf{t}) \ \dots \ \mathbf{x}_n(\mathbf{t})]$ . Our objective is to assign each set of data observations into clusters that can be efficiently represented by a low-dimensional subspace. This is equivalent to finding a partitioning  $\{\mathbf{X}^1(\mathbf{t}), \mathbf{X}^2(\mathbf{t}), \dots, \mathbf{X}^P(\mathbf{t})\}$  of  $[n]$  observations, where  $P$  is the total number of clusters

underlying each data modality indexed by  $\mathbf{p}$ . Furthermore, each linear subspace can be described as  $\mathcal{S}^{\mathbf{p}}(\mathbf{t}) \subset \mathbb{R}^m$  with  $\dim \mathcal{S}^{\mathbf{p}}(\mathbf{t}) \ll m$ .

We will exploit the self-expressive property presented in [1, 26], which entails that each data observation  $\mathbf{x}_i(\mathbf{t})$  can be represented as a linear combination of all features from the same subspace  $\mathcal{S}(\mathbf{x}_i(\mathbf{t}))$  as follows,

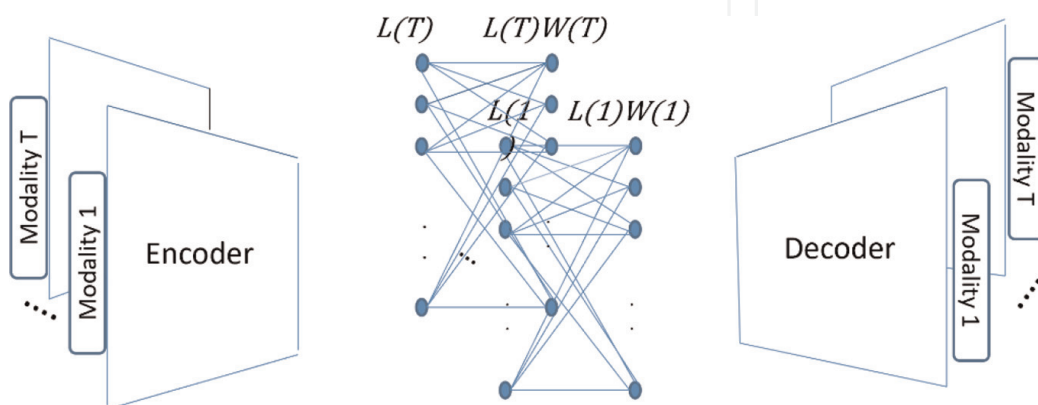
$$\mathbf{x}_i(\mathbf{t}) = \sum_{i \neq j, \mathbf{x}_j(\mathbf{t}) \in \mathcal{S}(\mathbf{x}_i(\mathbf{t}))} \mathbf{w}_{ij}(\mathbf{t}) \mathbf{x}_j(\mathbf{t}). \quad (1)$$

If we stack all the data points  $\mathbf{x}_i(\mathbf{t})$  into columns of the data matrix  $\mathbf{X}(\mathbf{t})$ . The self-expressive property can be written in a matrix form as follows,

$$\mathbf{X}(\mathbf{t}) = \mathbf{X}(\mathbf{t}) \mathbf{W}(\mathbf{t}) \text{ s.t. } \mathbf{W}_{ii} = \mathbf{0}. \quad (2)$$

The important information about the relations among data samples is then recorded in the self-representation coefficient matrix  $\mathbf{W}(\mathbf{t})$ . Under a suitable arrangement/permutation of the data realizations, the sparse coefficient matrix  $\mathbf{W}(\mathbf{t})$  is an  $n \times n$  block-diagonal matrix with zero diagonals provided that each sample is represented by other samples only from the same subspace. More precisely,  $\mathbf{W}_{ij}(\mathbf{t}) = \mathbf{0}$  whenever the indexes  $i, j$  correspond to samples from different subspaces. As a result, the majority of the elements in  $\mathbf{W}$  are equal to zero. A diagram showing our algorithm is depicted in **Figure 1**.

Our algorithm consists of three main stages; the first stage is the encoder which encodes the input modalities into a latent space. The encoder consists of  $T$  parallel CNNs, where  $T$  is the number of data modalities. Each modality data is fed into one network, and the output of each network represents the modality data projection into its corresponding hidden/latent space. The second component of the auto encoder is  $T$  self-expressive layers, the goal of which is to enforce the self-expressive property among the data observations of each data modality. Each self-expressive layer is a fully connected layer which independently operates on the output of each encoder. The last stage is the decoder which reconstructs input data from the self-expressive layers' output. The objective function sought through this approximation network is reflected in Eq. (5). The group sparsity introduced in [23] requires the minimization of the group norm of matrices  $\mathbf{W}(\mathbf{t})$ , which in turn, entails a smaller angle between the different spaces across all modalities, thus promoting the goal of obtaining a common latent space. Note that minimizing group norm provides a group sparse



**Figure 1.**  
Deep robust group subspace clustering diagram.

solution along data modalities. If we in addition, constrain the coefficient matrices corresponding to each data modality to commute, therefore, we ensure their sharing the same eigen vectors. The idea of commutation has been used in [27–29]. We define  $\Omega = \{\mathbf{W}(t)\}_{t=1}^T$ , where  $\mathbf{W}(t) = [w_{kj}(t)]_{k,j}$  and the group l-norm  $\|\Omega\|_{1,2}$  as:

$$\|\Omega\|_{1,2} = \sum_{k,j} \sqrt{\sum_{t=1}^T w_{kj}^2(t)}. \quad (3)$$

We also define  $[\mathbf{W}(t_1), \mathbf{W}(t_2)]$  as,

$$[\mathbf{W}(t_1), \mathbf{W}(t_2)] = \mathbf{W}(t_1)\mathbf{W}(t_2) - \mathbf{W}(t_2)\mathbf{W}(t_1), \mathbf{W}(t_1) = 0 \quad (4)$$

The loss function is then rewritten as,

$$\begin{aligned} \min_{\mathbf{W}(t)/w_{kk}(t)=0} \sum_{t_1, t_2}^T \|\mathbf{W}(t_1)\mathbf{W}(t_2)\|^2 + \|\Omega\|_{1,2} \\ + \frac{\gamma}{2} \sum_{t=1}^T \|\mathbf{X}(t) - \mathbf{X}_r(t)\|_F^2 \\ + \rho \sum_{t=1}^T \|\mathbf{W}(t)\|_1 + \frac{\mu}{2} \sum_{t=1}^T \|\mathbf{L}(t) - \mathbf{L}(t)\mathbf{W}(t)\|_F^2 \end{aligned} \quad (5)$$

where  $\mathbf{X}_r(t)$  represent the reconstructed data corresponding to modality  $t$ , and  $\mathbf{L}(t)$  is the output of the  $t^{\text{th}}$  encoder with input  $\mathbf{X}(t)$ .  $\mathbf{W}(t)$  is the sparse weight function that ties the data observation for modality  $t$ . Solving DRoGSuRe in Tensorflow and using the adaptive momentum based gradient descent method (ADAM) [30] results in minimizing the loss function. For each data modality, the weights of the encoder, the self-expressive layer and the decoder are individually calculated, however, fine-tuning the weights is based on the loss function, which is a function of the group norm and the pairwise product difference between sparse coefficient matrices.  $\|\cdot\|_1$  denotes the  $l_1$  norm, i.e., the sum of absolute values of the argument. The Lagrangian objective functional may be rewritten as,

$$\begin{aligned} L(\mathbf{W}(t)) = \sum_{t_1, t_2}^T \|\mathbf{W}(t_1)\mathbf{W}(t_2)\|^2 + \|\Omega\|_{1,2} \\ + \rho \sum_{t=1}^T \|\mathbf{W}(t)\|_1 + \frac{\gamma}{2} \sum_{t=1}^T \|\mathbf{X}(t) - \mathbf{X}_r(t)\|_F^2 \\ + \sum_{t=1}^T \frac{\mu}{2} \|\mathbf{L}(t)\mathbf{W}(t) - \mathbf{L}(t)\|_F^2 \\ + \sum_{t=1}^T \langle \mathbf{L}(t)\mathbf{W}(t) - \mathbf{L}(t), \mathbf{Y}(t) \rangle \end{aligned} \quad (6)$$

Assume  $\hat{\mathbf{W}}(t) = \mathbf{I} - \mathbf{W}(t)$ , we update  $\mathbf{W}(t)$  as follows,

$$\begin{aligned}
 \mathbf{W}_{k+1}(t) = \arg \min_{\mathbf{W}(t)} & \sum_{t_1, t_2}^T \|\mathbf{W}(t_1)\mathbf{W}(t_2)\|^2 + \|\mathbf{Q}\|_{1,2} + \rho\|\mathbf{W}(t)\|_1 + \\
 & \langle \mathbf{L}_{k+1}(t)\mathbf{W}(t) - \mathbf{L}_{k+1}(t), \mathbf{Y}_k(t) \rangle \\
 & > + \frac{\mu_k}{2} \|\mathbf{L}_{k+1}(t)\mathbf{W}(t) - \mathbf{L}_{k+1}(t)\|_F^2
 \end{aligned} \tag{7}$$

Similar to [4], we utilize linearized ADMM [31] to approximate the minimum of Eq. (7) since the algorithmic solution is complicated and yields a non-convex optimization functional. It has been shown that linearized ADMM is very effective for  $l_1$  minimization problems and the augmented Lagrange multiplier (ALM) method can take care of the non-convexity of the problem [32, 33]. Therefore, utilizing an appropriate augmented Lagrange multiplier  $\mu_k$ , we can compute the global optimizer by solving the dual problem. The solution to Eq. (7) can be approximated, using linearized soft thresholding, as follows,

$$\begin{aligned}
 \mathbf{W}_k^+(t) = \text{prox}_{\frac{\rho}{\mu_1}}(\mathbf{W}_k(t)) & + \frac{\mathbf{L}_{k+1}^T \left( \mathbf{L}_{k+1} \hat{\mathbf{W}}_k(t) - \frac{\mathbf{Y}_k(t)}{\mu_k} \right)}{\eta_1} \\
 & + \sum_{t_1, t_2=1, t_1 \neq t_2}^T \{ (\mathbf{W}_k(t_1)\mathbf{W}_k(t_2) \\
 & - \mathbf{W}_k(t_2)\mathbf{W}_k(t_1))\mathbf{W}_k^T(t_2) + \mathbf{W}_k(m)(\mathbf{W}_k(t_1)\mathbf{W}_k(t_2) \\
 & - \mathbf{W}_k(t_2)\mathbf{W}_k(t_1)) \} \\
 \mathbf{W}_{k+1}(t) = \gamma_{\frac{\rho}{\mu_2}}(\mathbf{W}_k^+(t)) & \tag{8} \\
 & \tag{9}
 \end{aligned}$$

where  $\eta_1 \geq \|\mathbf{L}\|_2^2$ . We alternatively update  $\mathbf{L}(t)$  as,

$$\mathbf{L}_{k+1}(t) = \mathbf{L}_k(t) + \mu_k \left( \mathbf{L}_k(t) \hat{\mathbf{W}}_{k+1}(t) - \frac{\mathbf{Y}_k(t)}{\mu_k} \right) \hat{\mathbf{W}}_{k+1}^T(t). \tag{10}$$

where  $\text{prox}_{\beta}(A_{i,j}(t)) = A_{i,j}(t) * \frac{\max\left\{\left(\sqrt{\sum_{t=1}^T A_{i,j}(t)^2} - \beta\right), 0\right\}}{\sqrt{\sum_{t=1}^T A_{i,j}(t)^2}}$  and

$\gamma_{\tau}(B_{i,j}) = \text{sign}(B_{i,j}) * \max\{|B_{i,j}| - \tau, 0\}$ . The Lagrange multipliers are updated as follows,

$$\mathbf{Y}_{k+1}(t) = \mathbf{Y}_k(t) + \mu_k(\mathbf{L}_{k+1}(t)\mathbf{W}_{k+1}(t) - \mathbf{L}_{k+1}(t)) \tag{11}$$

$$\mu_{k+1} = \epsilon\mu_k \tag{12}$$

After computing the gradient of the loss function, the weights of each multi-layer network, that corresponds to one modality, are updated while other modalities' networks are fixed. In other words, after constructing the data during the forward pass, the loss function determines the updates that back-propagates through each layer. The encoder of the first modality is updated, afterwards, the self-expressive layer of that modality gets updated and finally the decoder.

Since the weights corresponding to each modality are dependent on other modalities, we update each part of the network corresponding to each modality with the assumption that all other networks' components corresponding to other modalities are fixed. The resulting sparse coefficient matrices  $\mathbf{W}(\mathbf{t})$ 's, for  $\mathbf{t} = 1, 2, \dots, T$  are then integrated as follows,

$$\mathbf{W}_{Total} = \sum_{t=1}^T \mathbf{W}(t) \quad (13)$$

Integrating the sparse coefficient matrices helps reinforcing the relation between data points that exist in all data modalities, thus establishing a cross-sensor consistency. Furthermore, adding the sparse coefficient matrices reduces the noise variance introduced by the outliers. A similar approach was introduced in [34] for Social Networks community detection, where an aggregation of multi-layer adjacency matrices was proved to provide a better Signal to Noise ratio, and ultimately better performance. To proceed with distinguishing the various classes in unsupervised manner, we construct the affinity matrix as follows,

$$\mathbf{A} = \mathbf{W}_{Total} + \mathbf{W}_{Total}^T \quad (14)$$

where  $\mathbf{A} \in \mathbb{R}^{n \times n}$ . We subsequently use the spectral clustering method [35] to retrieve the clusters in the data using the above affinity matrix as input.

## 2.2 Theoretical discussion

In order to justify the multiple banks of self-expressive layers, we assume that each modality  $\mathbf{X}(\mathbf{t})$  may be expressed as a private information contribution  $\mathbf{X}_p(\mathbf{t})$  and a shared information  $\mathbf{X}_s(\mathbf{t})$  such that,

$$\mathbf{X}(\mathbf{t}) = \mathbf{X}_s(\mathbf{t}) + \mathbf{X}_p(\mathbf{t}) \quad (15)$$

The shared information can be represented as follows,

$$\mathbf{X}_s(\mathbf{t}) = \sum_{t=1}^T F(\mathbf{W}(t)(\Pi_s \mathbf{X}(t))) \quad (16)$$

where  $\Pi_s = \cap_{t=1, \dots, T} \Pi_s^t$ .  $\mathbf{X}_s(\mathbf{t})$  and  $\mathbf{X}_p(\mathbf{t})$  are distinct and will hence lie in different subspaces, which will hence be mapped to different components in  $\mathbf{W}(\mathbf{t})$ . Similarly for the subspaces spanned by  $\mathbf{X}_p(\mathbf{t}_i)$  and  $\mathbf{X}_p(\mathbf{t}_j)$ ,  $i \neq j$ , the corresponding components of  $\mathbf{W}(\mathbf{t}_i)$  and  $\mathbf{W}(\mathbf{t}_j)$  will almost surely not coincide. On the other hand, the components of  $\mathbf{W}(\mathbf{t}_i)$  and  $\mathbf{W}(\mathbf{t}_j)$  corresponding to  $\mathbf{X}_s(\mathbf{t}_i)$  and  $\mathbf{X}_s(\mathbf{t}_j)$  will almost surely coincide, thus justifying the construction of a layered  $\mathbf{W}_{Total}$ , and thereby improving the SNR. In addition, the decoder will help protect and maintain the private information corresponding to each modality  $\mathbf{X}_p(\mathbf{t})$  by ensuring that data can be reconstructed again from the latent space with minimal loss. In the following, we will elaborate more on how aggregating affinity matrices should impact the overall clustering performance. The idea of aggregating affinity



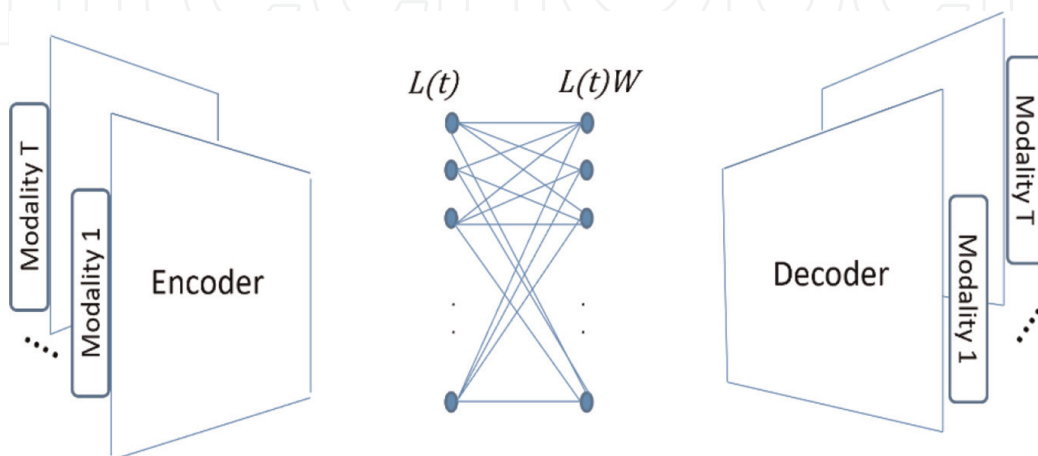
matrices is not new, in fact, it has been used extensively in clustering and community detection field. For example, in [36], the authors proposed a method that combines the self-similarity matrices of the eigenvectors after applying a Singular Value Decomposition on clusters. In [37], they proposed merging the information provided by the multiple modalities by combining the characteristics of individual graph layers using tools from subspace analysis on a Grassmann manifold. In [38], they propose a multilayer spectral graph clustering (SGC) framework that performs convex layer aggregation.

**Proposition:** The persistent differential scaling of  $m$ -modal Group Robust Subspace Clustering Fusion yields an order  $m$ -improvement resilience over the singly differential scaling fusion.

The proof of the proposition can be found in Appendix A. We basically show that by perturbing one or more data modalities, our proposed approach introduces less error to the overall affinity matrix as compared to DMSC. Hence, preserving the performance and yielding a graceful degradation of the clustering accuracy as an increasing number of modalities get corrupted by noise.

### 3. Affinity fusion deep multimodal subspace clustering

For completeness, we provide a brief overview of the Deep Multimodal Subspace Clustering algorithm which was proposed in [4]. As noted earlier for DRoGSuRe and similarly for Affinity Fusion Deep Multimodal Subspace clustering (AFDMSC), the network is composed of three main parts: a multimodal encoder, a self-expressive layer, and a multimodal decoder. The output of the encoder contributes to a common latent space for all modalities. The self-expressiveness property applied through a fully connected layer between the encoder and the decoder results in one common set of weights for all the data sensing modalities. This marks a divergence in defining the latent space with DRoGSuRe. Our proposed approach, as a result, safeguards the private information  $\mathbf{X}_p(t); t = 1, \dots, T$  individually for each of the sensors, i.e., dedicating more degrees of freedom for each of the sensors. This contrasts with AFDMSC. The reconstruction of the input data by the decoder, can yield the following loss function to secure the proper training of the self-expressive network,



**Figure 2.** Deep multimodal subspace clustering diagram.

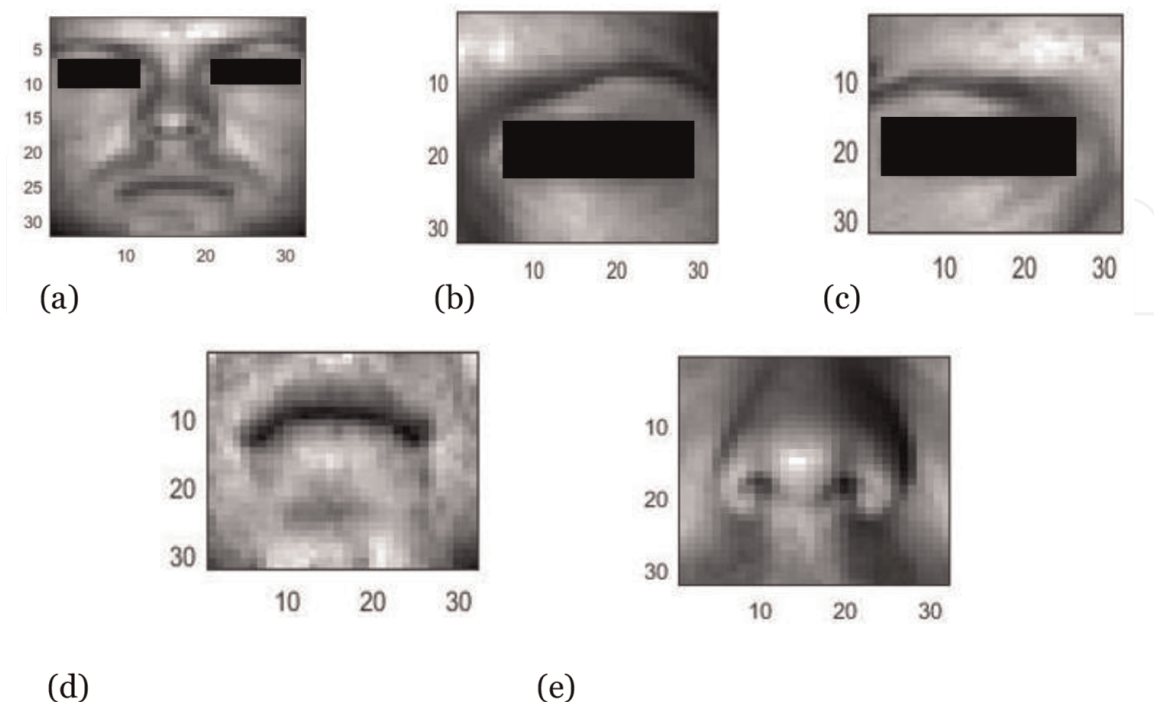
$$\min_{\mathbf{W} | w_{kk}=0} \|\mathbf{W}\|_2 + \frac{\gamma}{2} \sum_{t=1}^T \|\mathbf{X}(t) - \mathbf{X}_r(t)\|_F^2 + \frac{\mu}{2} \sum_{t=1}^T \|\mathbf{L}(t) - \mathbf{L}(t)\mathbf{W}\|_F^2, \quad (17)$$

where  $\mathbf{W}$  represents the parameters of the self expressive layer,  $\mathbf{X}(t)$  is the input to the encoder,  $\mathbf{X}_r(t)$  denote the output of the decoder and  $\mathbf{L}(t)$  denotes the output of the encoder.  $\mu$  and  $\gamma$  are regularization parameters. An overview for the DMSC approach is illustrated in **Figure 2**.

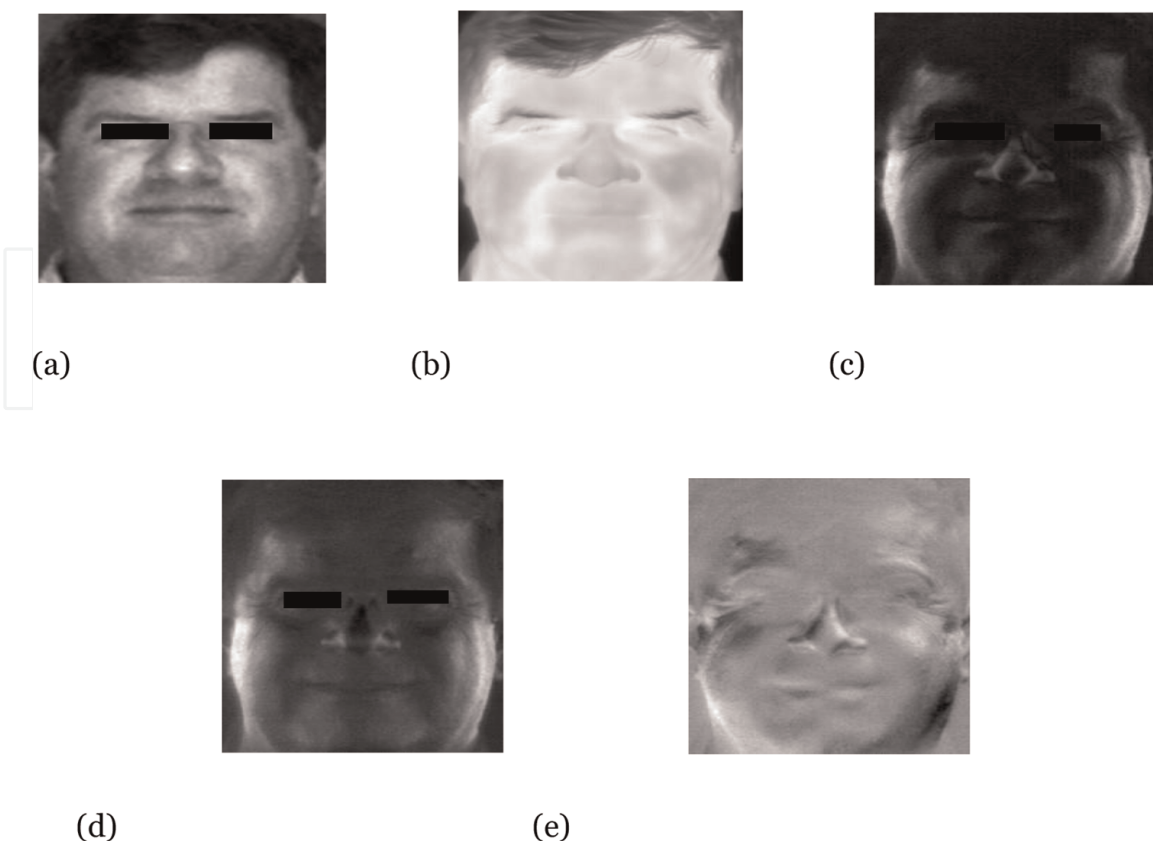
## 4. Experimental results

### 4.1 Dataset description

We will evaluate our approach on two different datasets. The first dataset we will use is the Extended Yale-B dataset [39]. The same dataset has been used extensively in subspace clustering as in [1, 40]. The dataset is composed of 64 frontal images of 38 individuals under different illumination conditions. In this work, we will use the augmented data used in [4], where facial components such as left eye, right eye, nose and mouth have been cropped to represent four additional modalities. Images corresponding to each modality have been cropped to a size of  $32 \times 32$ . A sample image for each modality is shown in **Figure 3**. The second validation dataset we use is the ARL polarimetric face dataset [41]. This consists of facial images for 60 individuals in the visible domain and in four different polarimetric states. All the images are spatially aligned for each subject. We have also resized the images to  $32 \times 32$  pixels. Sample images from this dataset are shown in **Figure 4**.



**Figure 3.** Sample images from the augmented extended Yale-B Dataset. (a) Face. (b) Left eye. (c) Right eye. (d) Mouth. (e) Nose.



**Figure 4.** Sample images from the ARL polarimetric dataset. (a) Visible. (b) DoLP. (c) So. (d) S1. (e) S2.

## 4.2 Network structure

In the following, we will elaborate on how we construct the neural network for each dataset. Similar to [4], we implemented DRoGSuRe with Tensorflow and used the adaptive momentum based gradient descent method (ADAM) [30] to minimize the loss function in Eq. (5) with a learning rate of  $10^{-3}$ .

In case of ARL dataset, we have five data modalities and will therefore have 5 different encoders, self-expressive layers and decoders. Each encoder is composed of three neural layers. The first layer consists of 5 convolutional filters of kernel size 3. The second layer has 7 filters of kernel size 1. The last layer has 15 filters with kernel size equals 1.

For EYB dataset, we also have five data modalities, therefore, we have 5 different encoders, self-expressive layers and decoders. Each encoder is composed of three neural layers. The first layer consists of 10 convolutional filters of kernel size 5. The second layer has 20 filters of kernel size 3. The last layer has 30 filters of kernel size 3.

## 4.3 Noiseless results

In the following, we compare the performance of our approach versus the DMSC approach when learning the union of subspaces structure of noise-free data. First, we divide each dataset into training and validation sets to be able to classify a newly observed dataset, using the structure learned through the current unlabeled data. The ARL expression dataset used for training consists of 2160 images per modality. The validation baseline images include 720 images total per modality. For the EYB, we

	Learning	Validation
DMSC	97.59%	98.33%
DRoGSuRe	100%	100%

**Table 1.**  
 Performance comparison for ARL dataset.

	Learning	Validation
DMSC	98.82%	98.89%
DRoGSuRe	98.42%	98.76%

**Table 2.**  
 Performance comparison for EYB dataset.

randomly selected 1520 images per modality for training and 904 images for validation. The sparse solution  $\mathbf{W}(\mathbf{t})$  corresponding to each data modality, provides important information about the relations among data points, which may be used to split data into individual clusters residing in a common subspace. Observations from each object can be seen as data points spanning one subspace. Interpreting the subspace-based affinities based on  $\mathbf{W}(\mathbf{t})$  as a layered set of networks, we proceed to carry out what amounts to modality fusion. The  $\mathbf{T}$  sparse matrices are added to produce one sparse matrix for both modalities,  $\mathbf{W}_{Total}$ , thereby improving performance. Observations associated with one object/individual are clustered as one subspace where the contribution of each sensor is embedded in the entries of the  $\mathbf{W}_{Total}$  matrix. For clustering by  $\mathbf{W}_{Total}$ , we apply spectral clustering.

After learning the structure of the data clusters, we validate our results on the validation set. We extract the principal components (eigen vectors of the covariance matrix) of each cluster in the original (training) dataset, to act as a representative subspace of its corresponding class. We subsequently project each new test point onto the subspace corresponding to each cluster, spanned by its principal components. The  $l_2$  norm of the projection is then computed, and the class with the largest norm is selected to be the class of this test point. For DRoGSuRe, we use the coefficient matrix  $\mathbf{W}_{Total}$  in Eq. (13) to cluster the test data points coming from all data modalities. We compare the clustering output labels with the ground truth for each dataset. The results for ARL and EYB datasets are depicted in **Tables 1** and **2** respectively. From the results, it is clear that DRoGSuRE technique for the fused data remarkably outperforms DMSC in case of ARL dataset. The reason behind the significant improvement is the layered structure of our proposed approach that constructs the latent space in a way that safeguards the individual sensor private information which hence dedicates more degrees of freedom to each of the sensors. In addition, the ARL dataset structure offers modalities that are different in nature and individually provides new information in contrast to the EYB dataset. However, in case of EYB dataset and in the noiseless case, DMSC performed better than DRoGSuRe.

#### 4.4 Noise training with single and multiple modalities

In the following, we test the robustness of our approach in the case of noisy learning. We distort one modality at a time by shuffling the pixels of all images in that

	DMSC L	DMSC V	DRoGSuRe L	DRoGSuRE V
Mod 0	87.17%	86.67%	95.37%	95%
Mod 1	91.67%	90%	98.29%	98.33%
Mod 2	92.77%	92.78%	99.17%	99.44%
Mod 3	90.55%	90.57%	99.31%	99.44%
Mod 4	92.78%	91.11%	96.44%	96.67%

**Table 3.**  
ARL dataset: Distorting one modality.

	DMSC L	DMSC V	DRoGSuRe L	DRoGSuRE V
Mod 0 & 1	82.22%	82.78%	92.27%	94.58%
Mod 1 & 2	91.11%	91.11%	97.22%	97.36%
Mod 0 & 3	85.51%	82.56%	93.01%	95.42%
Mod 1 & 4	91.67%	89.44%	97.22%	97.36%
Mod 2 & 3	90%	89.72%	97.69%	97.78%

**Table 4.**  
ARL dataset: Distorting two modalities.

	DMSC L	DMSC V	DRoGSuRe L	DRoGSuRE V
Mod 0	87.96%	88.5%	93.29%	94.69%
Mod 1	91.84%	91.15%	95.79%	97.46%
Mod 2	89.01%	88.72%	98.03%	97.57%
Mod 3	92.69%	91.81%	95.59%	96.68%
Mod 4	91.45%	91.59%	97.17%	97.35%

**Table 5.**  
EYB dataset: Distorting one modality.

modality during the training phase. By doing so, we are perturbing the structure of the sparse coefficient matrix associated with that modality, thus impacting the overall  $W$  matrix for both DRoGSuRe and DMSC. Testing with clean data, i.e., no distortion, demonstrates the impact of perturbing the training and hence performing an

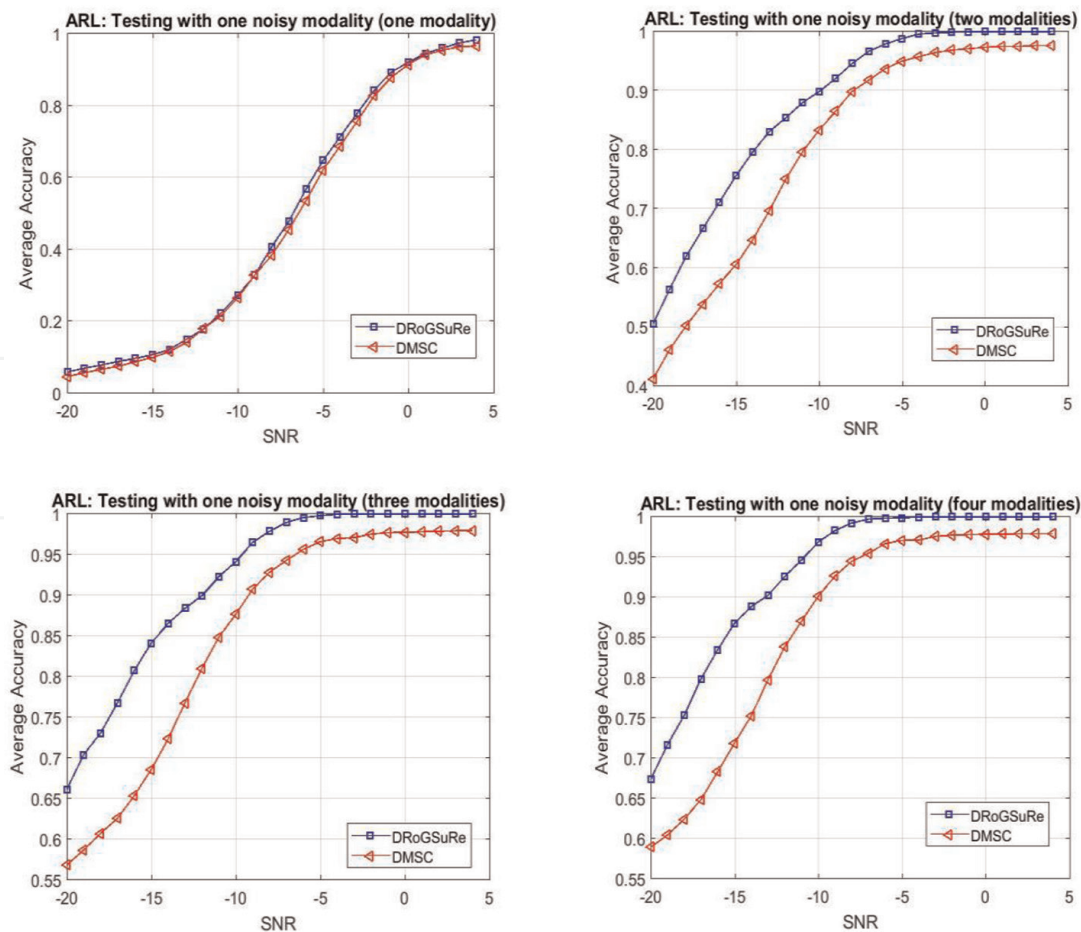
	DMSC L	DMSC V	DRoGSuRe L	DRoGSuRE V
Mod 0 & 2	86.64%	85.18%	96.84%	96.13%
Mod 0 & 4	87.83%	89.16%	94.54%	95.8%
Mod 1 & 4	86.38%	86.06%	94.21%	95.8%
Mod 2 & 3	88.22%	84.96%	91.58%	93.92%
Mod 3 & 4	88.03%	86.28%	94.08%	95.35%

**Table 6.**  
EYB dataset: Distorting two modalities.

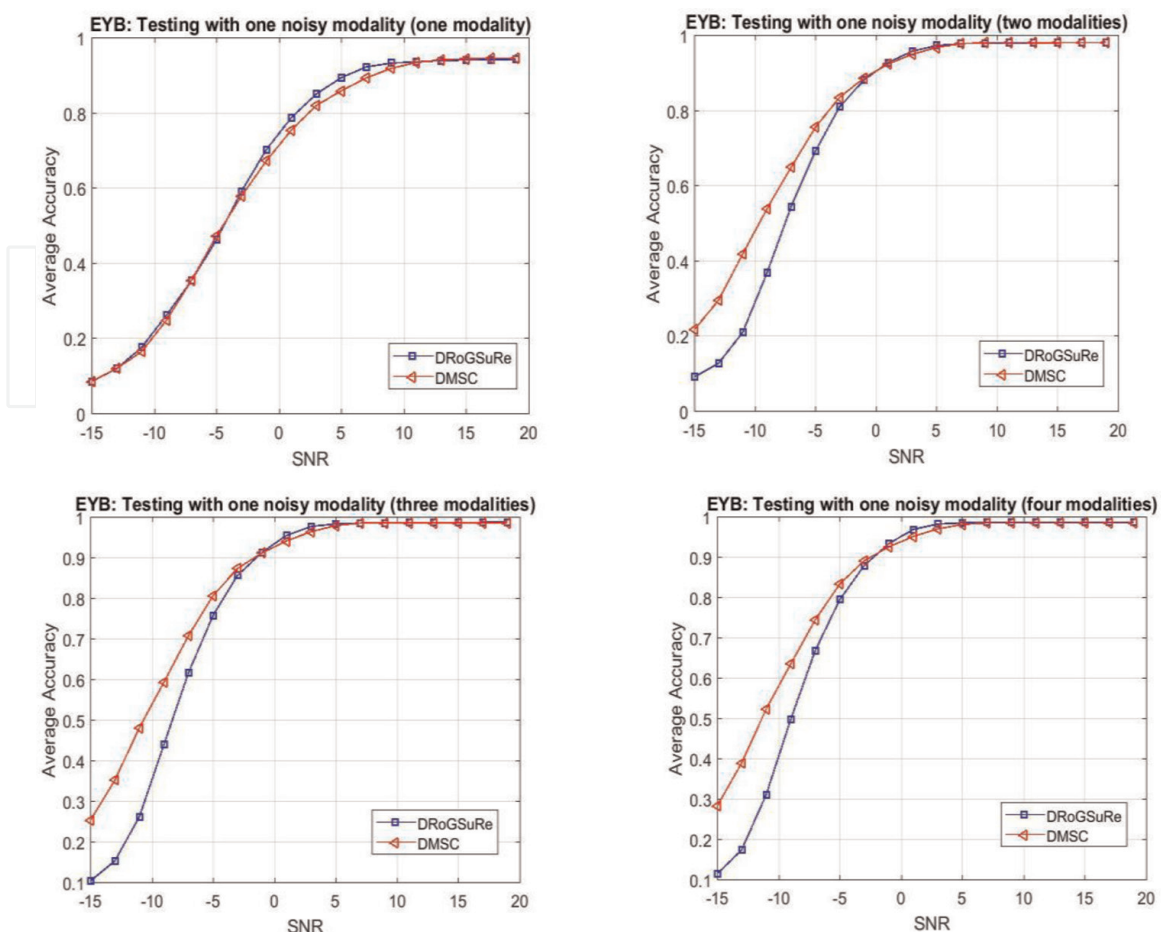
inadequate training, e.g., insufficient data or non-convergence. This can also be considered as augmenting the training data with new information or a new view for one modality which might not necessarily contained in the testing or the validation data. Moreover, we repeat the same experiment with the distortion of two modalities before learning the sparse coefficient matrices for both DMSC and DRoGSuRe. The results for the ARL dataset are depicted in **Tables 3** and **4**, while results for the EYB dataset are shown in **Tables 5** and **6**. For ARL dataset, we refer to Visible, S0, S1, S2 and DoLP as Mod 0, 1, 2, 3 and 4 respectively. For EYB Dataset, we refer to Face, left eye, nose, mouth and right eye as mod 0, 1, 2, 3, and 4. We refer to each modality as Mod, where L denotes learning and V denotes validation results. From the results, it is clear that DRoGSuRe is showing a significant improvement in the clustering accuracy as compared to DMSC for both learning and validation set. The reason for that, is again, due to the fact that perturbing one or two modalities would have less impact on the overall performance for DRoGSuRe in comparison to DMSC.

#### 4.5 Testing with limited noisy testing data

In the following, we study the effect of using noiseless data for training while validating with noisy and missing data. We add Gaussian noise to one data modality in the validation set and vary the SNR by varying the noise variance. We subsequently assume that we only have one modality available at testing. Then, we keep increasing the number of available noiseless data modalities beside the noisy modality. We



**Figure 5.**  
 ARL noiseless training and validating on limited noisy data.

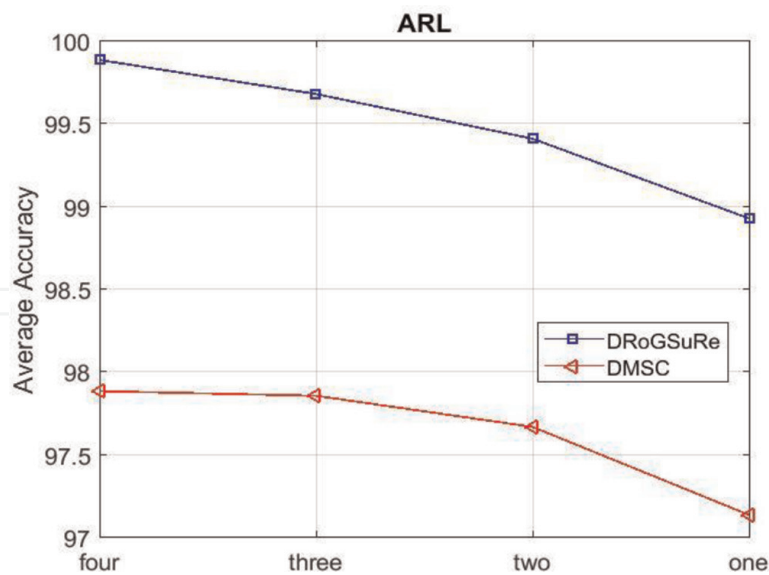


**Figure 6.**  
EYB noiseless training and validating on limited noisy data.

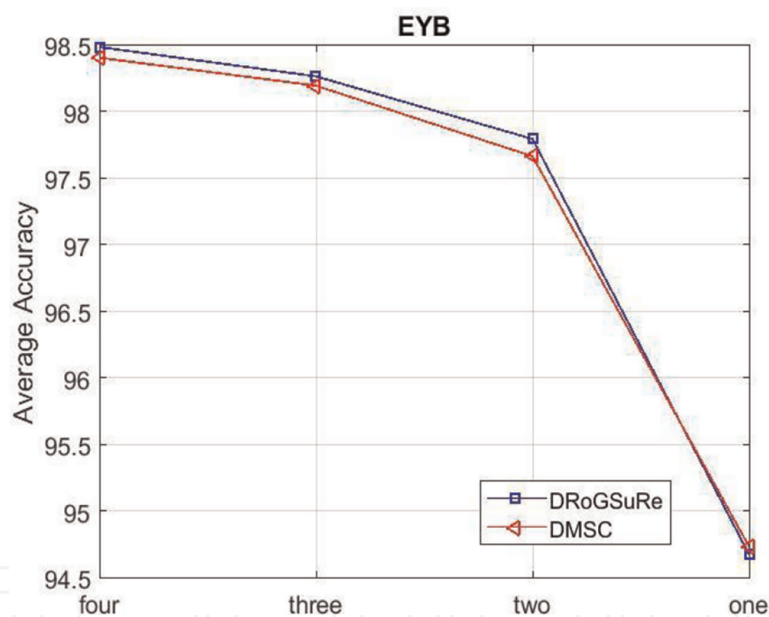
average the results considering all different combinations of data modalities for ARL and EYB datasets. The results are depicted in **Figures 5** and **6** respectively. For the ARL dataset, we note the increasing gap between DMSC and DRoGSuRe as we augment the sensing capacity with noise-free modalities. On the other hand, for the EYB dataset and at lower SNR, the performance of DRoGSuRe is slightly worse than DMSC which might be explained by the results in **Table 2**; as the training accuracy for DMSC is slightly better than DRoGSuRe in the case of clean training. However, at higher SNR, the performance of the two approaches is very close.

#### 4.6 Missing modalities during testing

In the following, we evaluate the performance of DRoGSuRe and DMSC in case of missing data modalities during testing. It is not uncommon to have one or more sensors that might be silent during testing, thus justifying this experiment for further assessment. We try different combinations of available modalities during testing, and we average the clustering accuracy for each trial. Results are depicted in **Figures 7** and **8** for ARL and EYB data respectively. Again, we notice a significant improvement for DRoGSuRe over DMSC for ARL Dataset. For EYB dataset, there is a slight improvement for DRoGSuRe over DMSC. The reason behind the slight improvement is because our approach introduces less error to the overall affinity matrix as compared



**Figure 7.**  
*Missing modalities during testing for ARL dataset.*



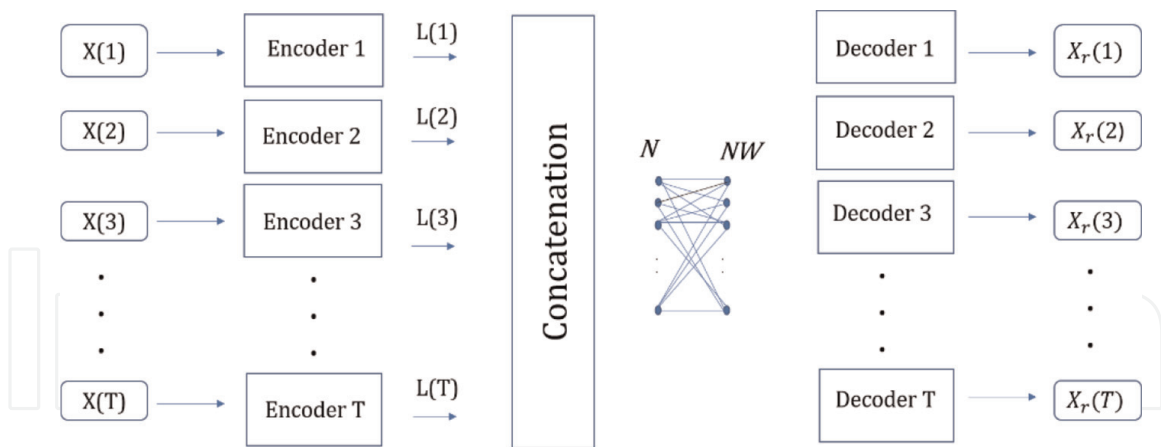
**Figure 8.**  
*Missing modalities during testing for EYB dataset.*

to DMSC. Hence, preserving the performance and yielding a graceful degradation of the accuracy although DMSC was the state of the art for the EYB dataset.

## 5. Feature concatenation

Here we propose a rationale along with an alternative solution for enhancing the performance for EYB multi-modal data. Due to the specific structure of the EYB multi-modal data, the concatenation of the features corresponding to each modality is a reasonable alternative. By doing so, we are adjoining together the features representing each part of the face. Since the four modalities correspond to non-overlapping partitions of the face, the feature set corresponding to each partition will





**Figure 9.**  
CNN Concatenation Network.

solely provide complementing information. A similar idea is proposed in [4] and is referred to as Late concatenation, where the multi-modal data is integrated in the last stage of the encoder. Their resulting decoder structure remains the same for either affinity fusion or late concatenation. This entails de-concatenating the multi-modal data prior to decoding it. Our proposed approach on the other hand, results in a self-expressive layer being driven by the concatenated features from the  $M$  encoder branches. Afterwards, we feed the self-expressive layer output to each branch of the decoder. The concatenated information results in a more efficient code for the data, thereby resulting in an overall parsimonious with a sparse structure of the decoder, results in a decoder composed of three neural layers. The first layer consists of 150 filters of kernel size 3. The second layer consists of 20 layers of kernel size 3. The third layer consists of 10 layers of kernel size 5. Our approach is illustrated in **Figure 9**. We optimize the weights of the auto-encoder as follows,

$$\min_{W|w_{kk}=0} \rho \|W\|_1 + \frac{\gamma}{2} \sum_{t=1}^T \|\mathbf{X}(t) - \mathbf{X}_r(t)\|_F^2 + \frac{\mu}{2} \sum_{t=1}^T \|N - NW\|_F^2, \quad (18)$$

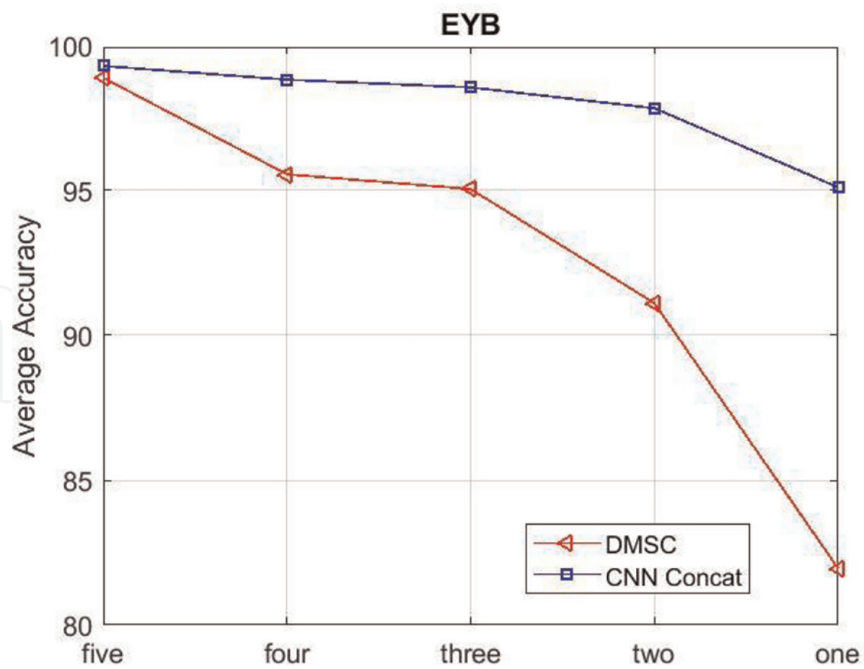
where  $N = [L(1) \| L(2) \| L(3) \| L(4) \| L(5)]$ .

We compared the performance of our proposed approach against the late concatenation approach in [4] and the results are depicted in **Table 7** for the EYB dataset.

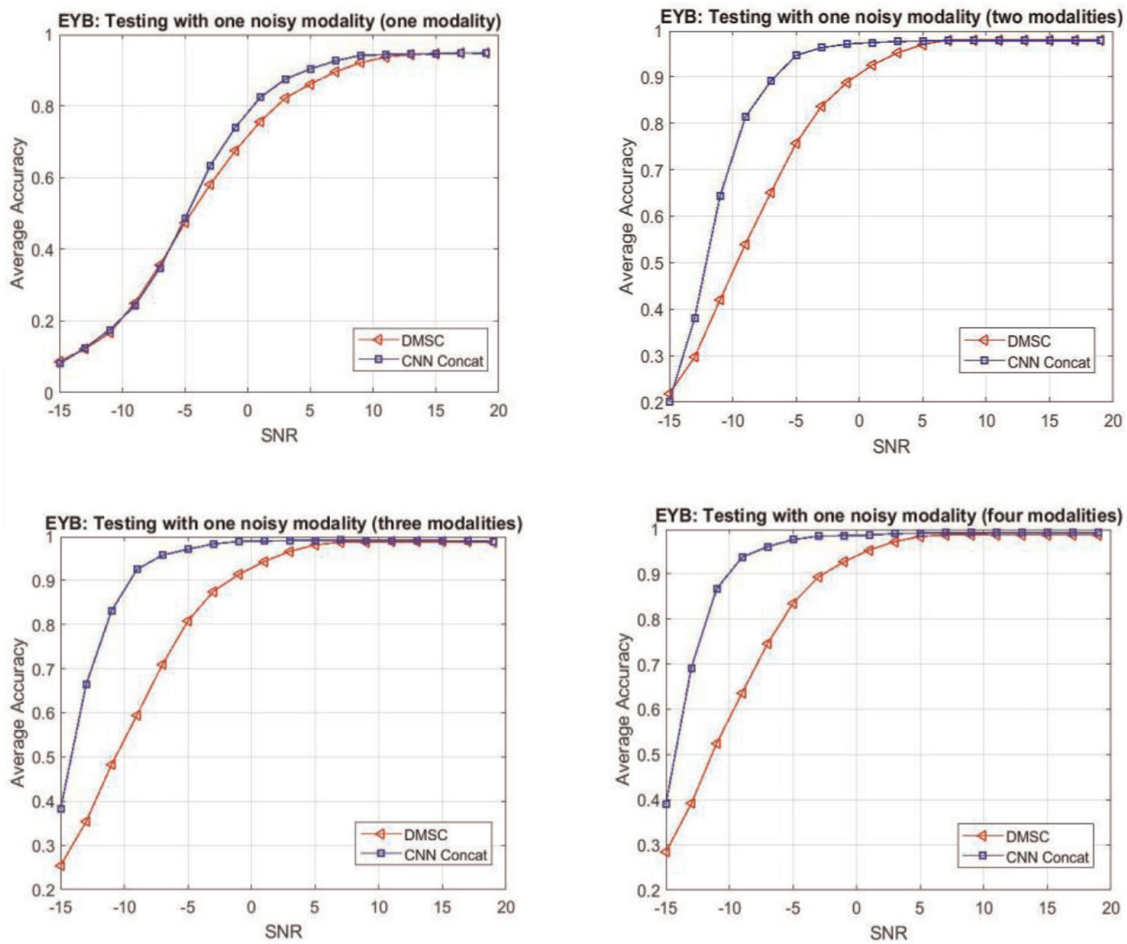
From the previous table, we can conclude that concatenating the features from the encoder and feeding the concatenated information to each decoder branch achieves a better performance for this type of multi-modal data structure. The reason behind this enhancement is the combination of efficient extraction of the basic features from the whole face and finer features from each part of the face. Promoting more efficiency as noted, this concatenation may also be intuitively viewed as adequate mosaicking, in which different patterns complement each other. In the following, we will show how

	Learning	Validation
DMSC Late Concatenation	95.66%	94.7%
CNN Concatenation Network	99.28%	99.3%

**Table 7.**  
Concatenation performance for EYB dataset.



**Figure 10.**  
 Missing modalities during testing for EYB dataset.



**Figure 11.**  
 EYB noiseless training and validating on limited noisy data.

	Learning	Validation
DMSC	97.59%	98.33%
DRoGSuRe	100%	100%
CNN Concat	99.44%	99.17

**Table 8.**  
*Concatenation performance for ARL dataset.*

our proposed approach performs in two cases: missing and noisy test data. The results of the new proposed approach, which we refer to as CNNs concatenation network, is compared to the state-of-the-art DMSC network [4]. We start by training the auto-encoder network using 75% of the data and then we test on the rest of the data. In **Figure 10**, we show how the performance degrades by decreasing the number of available modalities at testing from five to one. From the results, it is clear how the CNNs concatenation network outperforms the DMSC network. Additionally, we repeated the same experiment we performed in subsection 4.5. We train the network with noiseless data and then add Gaussian noise to one data modality at the testing. Additionally, we vary the number of available modalities at testing from one to four. The results are depicted in **Figure 11**. From the results, it is clear how the concatenated CNNs is more robust to noise than DMSC.

In addition, we have utilized the Concatenation network to perform object clustering on the ARL data. We compare the clustering performance of the concatenation network with both DMSC and DRoGSuRe. The results are depicted in **Table 8**. From the results, we conclude that DRoGSuRe still outperforms the other approaches for the ARL dataset. Although the number of parameters involved in training the DRoGSuRe network is higher than other approaches, since there are multiple self-expressive layers, however, DRoGSuRe is more robust to noise and limited data availability during testing.

## 6. Conclusion

In this paper, we proposed a deep multi-modal approach to fuse data through recovering the underlying subspaces of data observations from data corrupted by noise to scale to complex data scenarios. DRoGSuRe provides a natural way to fuse multi-modal data by employing the self-representation matrix as an embedding for each data modality. Experimental results show a significant improvement for DRoGSuRe over DMSC under different types of potential limitations and provides robustness with limited sensing modalities. We also proposed the concatenated CNNs model, which can work better for different multi-modal data structures.

## Acknowledgements

This work was in part supported by DOE-National Nuclear Security Administration through CNEC-NCSU under Award DE-NA0002576. The first author was also in part supported by DTRA.

## Conflict of interest

The authors declare no conflict of interest.

## A. Appendix. Parameter perturbation analysis

To theoretically compare our proposed variational scaling fusion approach DRoGSuRe to DMSC, we proceed by way of a first order perturbation analysis on the parameter set  $W^i$  of respectively either technique  $i = 1, 2$ . This will, in turn impact the associated affinity matrix  $A^i$ , which as we will later elaborate directly impacts the subspace clustering procedure which is central to the inference following the fusion procedure.

Adopting the original formulation for the first persistently differential scaling approach, namely that  $T$  modalities are jointly exploited, results in,  $X^1(t) = [x_1^1(t)x_2^1(t) \dots x_n^1(t)]$ , where  $x_k^1(t) \in \mathbb{R}^m$ ,  $t = 1, 2, \dots, T$  represents the  $k^{\text{th}}$  observation. The second approach only effectively uses only one subspace structure of the fused modalities  $X^2(t) = [x_1^1x_2^1 \dots x_n^1]$ .

A first order perturbation on the data may be due to noise or to a degradation of a given sensor, and results in a perturbation of the UoS parameters,

$$\tilde{W}_1^i = W_1^i + \delta^i \quad (\text{A1})$$

For the first method, each modality will have an associated subspace cluster parameter set  $\{W_t^1\}_{t=1, \dots, T}$ , with  $W_t^1 \in \mathbb{R}^{n \times n}$ . The overall parameter set for DRoGSuRe can then be written as,

$$\tilde{W}^1 = \tilde{W}_1^1 + W_2^1 + \dots + W_m^1 \quad (\text{A2})$$

Where the unperturbed overall sparse coefficient matrix is written as follows,  $W_{tot}^1 = W_1^1 + W_2^1 + \dots + W_m^1$ . A similar development follows for method 2, with the difference that the contributing modalities are fused a priori.

Proof. We first write the affinity matrix associated with DRoGSuRE as,

$$\tilde{A}^1 = \tilde{W}_{tot}^1 + \left(\tilde{W}_{tot}^1\right)^T \quad (\text{A3})$$

$$\tilde{A}^1 = \tilde{W}_1^1 + W_2^1 + \dots + W_m^1 + \left(\tilde{W}_1^1 + W_2^1 + \dots + W_m^1\right)^T \quad (\text{A4})$$

where the superscript  $T$  denotes transpose. This is equivalent to,

$$\tilde{A}^1 = \tilde{A}_1^1 + \sum_{i=2}^T A_i^1 \quad (\text{A5})$$

Where  $0 \leq \tilde{A}_1^1(i, j) \leq 1 + \delta^1$ . The unperturbed collective affinity matrix  $A^1$  can be similarly written  $A^1 = \sum_{i=1}^T A_i^1$  with the unity constraint on each entry of all matrices. We may also write the magnitude of the difference as,

$$|\mathbf{A}^1 - \tilde{\mathbf{A}}^1| = \delta^1 + (\delta^1)^T \quad (\text{A6})$$

Letting  $\Delta = \delta^1 + (\delta^1)^T \in \mathbb{R}^{n \times n}$ , and assuming  $\epsilon = \max_{i,j}[\Delta]_{i,j}$ , we can write,

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_F \leq n\epsilon \quad (\text{A7})$$

Given the  $\Delta$  matrix individual entry bounds, we conclude that,

$$0 \leq \epsilon \leq \frac{1}{t} \quad (\text{A8})$$

Since DMSC assumes having one sparse coefficient matrix  $\mathbf{W}$  for all data modalities, which is equivalent to only one subspace structure of the fused modalities  $\mathbf{X}^2(\mathbf{t}) = [\mathbf{x}_1^2 \dots \mathbf{x}_n^2]$ . Therefore, the UoS parameters will be perturbed by  $\delta^2$  as follows,

$$\tilde{\mathbf{W}}^2 = \mathbf{W}^2 + \delta^2 \quad (\text{A9})$$

The affinity matrix associated with DMSC can be written as follows,  $\tilde{\mathbf{A}}^2 = \tilde{\mathbf{W}}^2 + (\tilde{\mathbf{W}}^2)^T$ , which is equivalent to,

$$\tilde{\mathbf{A}}^2 = \mathbf{W}^2 + \delta^2 + (\mathbf{W}^2)^T + (\delta^2)^T \quad (\text{A10})$$

Similarly, the unperturbed affinity matrix will be as follows,

$$\mathbf{A}^2 = \mathbf{W}^2 + (\mathbf{W}^2)^T \quad (\text{A11})$$

From Eqs. (A10) and (A11), the magnitude of the difference can be written as follows,

$$|\mathbf{A}^2 - \tilde{\mathbf{A}}^2| = \delta^2 + (\delta^2)^T \quad (\text{A12})$$

Letting  $\gamma = \delta^2 + (\delta^2)^T \in \mathbb{R}^{n \times n}$ , i.e.,  $|\mathbf{A}^2 - \tilde{\mathbf{A}}^2| = \gamma$ , and assuming  $\Psi = \max_{i,j}[\gamma]_{i,j}$ , we can write  $\|\mathbf{A}^2 - \tilde{\mathbf{A}}^2\|_F \leq n\Psi$ . Given the  $\gamma$  matrix individual entry bounds, we conclude  $\mathbf{0} \leq \Psi \leq \mathbf{1}$ . If we only perturb one modality, knowing that  $\mathbf{0} \leq \mathbf{A}(i,j) \leq \mathbf{1}$ , therefore the error could lie between  $\mathbf{0} \leq \Psi \leq \mathbf{1}$ , which entails either creating a fake relation between two data points or erasing an existing relation.  $\epsilon$  and  $\Psi$  are random variables that do not have to follow a specific distribution, however, in any case  $E(\epsilon^2) \ll E(\Psi^2)$  and therefore  $SNR_{DRoGSuRe} \gg SNR_{DMSC}$ .

In light of the above two bounds, and the results of [42], where it is shown that the spectral clustering dependent on the respective projection operators  $\mathbf{P}_{\mathbf{W}^1}$  and  $\tilde{\mathbf{P}}_{\tilde{\mathbf{W}}^1}$  onto the vector subspaces spanned by the principal eigenvectors of  $\mathbf{W}_{tot}^1$  and  $\tilde{\mathbf{W}}_{tot}^1$  of may be written as,

$$\|\mathbf{P}_{\mathbf{W}^1} - \tilde{\mathbf{P}}_{\tilde{\mathbf{W}}^1}\|_F \leq \frac{\sqrt{2}}{\alpha^1} \|\mathbf{A}^1 - \tilde{\mathbf{A}}^1\|_F \quad (\text{A13})$$

where  $\alpha^1$  is the spectral gap between the  $k^{th}$  and  $(k + 1)^{st}$  eigen value of  $A^1$ ,  $|\lambda_k^1 - \lambda_{k+1}^1|$ . Similarly, for DMSC, the bound on the projection operators is,

$$\|P_{W^2} - \tilde{P}_{\tilde{W}^2}\|_F \leq \frac{\sqrt{2}}{\alpha^2} \|A^2 - \tilde{A}^2\|_F \quad (A14)$$

where  $\alpha^2 = |\lambda_k^1 - \lambda_{k+1}^1|$ . Since  $W_1^1, W_2^1, \dots, W_T^1$  happen to commute and if they happen to be diagonalizable, therefore, they share the same eigenvectors. As a result, the eigenvectors of  $W_1^1 + W_2^1 + \dots + W_T^1$  are also the same and the corresponding eigenvalue that is the sum of the corresponding eigenvalues of  $W_1^1, W_2^1, \dots$  and  $W_T^1$ . Therefore,  $\lambda_k^1 \gg \lambda_k^2$ . From all the above, we can conclude that smaller error yielding to better clustering, hence preserving the performance, yields the improvement by the T-factor noted in the proposition and shown in the two perturbation developments.

IntechOpen


### Author details

Sally Ghanem and Hamid Krim\*  
North Carolina State University, Raleigh, USA

\*Address all correspondence to: [ahk@ncsu.edu](mailto:ahk@ncsu.edu)

### IntechOpen

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Elhamifar E, Vidal R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013;**35**:2765-2781
- [2] Favaro P, Vidal R, Ravichandran A. A closed form solution to robust subspace estimation and clustering. In: *CVPR 2011*. Colorado springs, Colorado, USA: IEEE; 2011. pp. 1801-1807
- [3] Li CG, Vidal R. Structured sparse subspace clustering: A unified optimization framework. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA, USA: IEEE; 2015. pp. 277-286
- [4] Bian X, Panahi A, Krim H. Bi-sparsity pursuit: A paradigm for robust subspace recovery. *Signal Processing*. 2018;**152**: 148-159
- [5] Yang AY, Rao SR, Ma Y. Robust statistical estimation and segmentation of multiple subspaces. In: *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. New York, NY, USA: IEEE; 2006. p. 99
- [6] Hong W, Wright J, Huang K, Ma Y. Multiscale hybrid linear models for lossy image representation. *IEEE Transactions on Image Processing*. 2006;**15**:3655-3671
- [7] Ho J, Yang MH, Lim J, Lee KC, Kriegman D. Clustering appearances of objects under varying illumination conditions. In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003*. Proceedings. Madison, Wisconsin, USA: IEEE; 2003. p. I
- [8] Hellwich O, Wiedemann C. Object extraction from high-resolution multisensor image data. In: *Third International Conference Fusion of Earth Data*. France: Sophia Antipolis; 2000
- [9] Korona Z, Kokar MM. Model theory based fusion framework with application to multisensor target recognition. In: *1996 IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems (Cat. No. 96TH8242)*. Tokyo, Japan: IEEE; 1996. pp. 9-16
- [10] Ghanem S, Panahi A, Krim H, Kerekes RA, Mattingly J. Information subspace-based fusion for vehicle classification. In: *2018 26th European Signal Processing Conference (EUSIPCO)*. Rome, Italy: IEEE; 2018. pp. 1612-1616
- [11] Ghanem S, Roheda S, Krim H. Latent code-based fusion: A volterra neural network approach. 2021. arXiv preprint arXiv:2104.04829
- [12] Wang H, Skau E, Krim H, Cervone G. Fusing heterogeneous data: A case for remote sensing and social media. *IEEE Transactions on Geoscience and Remote Sensing*. 2018;**56**:6956-6968
- [13] Soong FK, Rosenberg AE. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1988;**36**:871-879
- [14] Xu L, Krzyzak A, Suen CY. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*. 1992;**22**: 418-435
- [15] Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng A. Multimodal deep learning.

- In: International Conference on Machine Learning (ICML). Bellevue, Washington, USA: International Machine Learning Society (IMLS). The conference. 2011. pp. 689-696
- [16] Ramachandram D, Taylor GW. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*. 2017;**34**: 96-108
- [17] Valada A, Oliveira GL, Brox T, Burgard W. Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In: International Symposium on Experimental Robotics. Tokyo, Japan: Springer; 2016. pp. 465-477
- [18] Roheda S, Riggan BS, Krim H, Dai L. Cross-modality distillation: A case for conditional generative adversarial networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Alberta, Canada: IEEE; 2018b. pp. 2926-2930
- [19] Roheda S, Krim H, Luo ZQ, Wu T. Decision level fusion: An event driven approach. In: 2018 26th European Signal Processing Conference (EUSIPCO). Rome, Italy: IEEE; 2018a. pp. 2598-2602
- [20] Zhu P, Hui B, Zhang C, Du D, Wen L, Hu Q. Multi-view deep subspace clustering networks. 2019. arXiv preprint arXiv:1908.01978.
- [21] Abavisani M, Patel VM. Deep multimodal subspace clustering networks. *IEEE Journal of Selected Topics in Signal Processing*. 2018;**12**: 1601-1614
- [22] Wang Q, Cheng J, Gao Q, Zhao G, Jiao L. Deep multi-view subspace clustering with unified and discriminative learning. *IEEE Transactions on Multimedia*. 2020;**23**: 3483-3493
- [23] Ghanem S, Panahi A, Krim H, Kerekes RA. Robust group subspace recovery: A new approach for multi-modality data fusion. *IEEE Sensors Journal*. 2020;**20**:12307-12316
- [24] Sun X, Cheng M, Min C, Jing L. Self-supervised deep multi-view subspace clustering. In: Asian Conference on Machine Learning, Nagoya, Japan: PMLR. 2019. pp. 1001-1016
- [25] Ji P, Zhang T, Li H, Salzmann M, Reid I. Deep subspace clustering networks. *Advances in Neural Information Processing Systems*. 2017;**30**:24-33
- [26] Bian X, Krim H. Bi-sparsity pursuit for robust subspace recovery. In: 2015 IEEE International Conference on Image Processing (ICIP). Québec city, Québec, Canada: IEEE; 2015. pp. 3535-3539
- [27] Roheda S, Krim H, Riggan BS. Robust multi-modal sensor fusion: An adversarial approach. *IEEE Sensors Journal*. 2020b;**21**:1885-1896
- [28] Roheda S, Krim H, Riggan BS. Commuting conditional GANS for multi-modal fusion. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE; 2020a. pp. 3197-3201
- [29] Roheda S, Krim H, Luo ZQ, Wu T. Event driven fusion. 2019. arXiv preprint arXiv:1904.11520
- [30] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014. arXiv preprint arXiv:1412.6980
- [31] Lin Z, Liu R, Su Z. Linearized alternating direction method with



adaptive penalty for low-rank representation. In: *Advances in Neural Information Processing Systems*. Granada, Spain: NIPS; 2011. pp. 612-620

[32] Rockafellar RT. Augmented Lagrange multiplier functions and duality in nonconvex programming. *SIAM Journal on Control*. 1974;**12**: 268-285

[33] Luenberger DG, Ye Y, et al. *Linear and Nonlinear Programming*. Vol. 2. Springer; 1984

[34] Taylor D, Shai S, Stanley N, Mucha PJ. Enhanced detectability of community structure in multilayer networks through layer aggregation. *Physical Review Letters*. 2016;**116**: 228301

[35] Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm. In: *Advances in Neural Information Processing Systems*. Vancouver, British Columbia, Canada: NIPS; 2002. pp. 849-856

[36] Gabriel HH, Spiliopoulou M, Nanopoulos A. Eigenvector-based clustering using aggregated similarity matrices. In: *Proceedings of the 2010 ACM Symposium on Applied Computing*. Switzerland: Sierre; 2010. pp. 1083-1087

[37] Dong X, Frossard P, Vandergheynst P, Nefedov N. Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds. *IEEE Transactions on Signal Processing*. 2013; **62**:905-918

[38] Chen PY, Hero AO. Multilayer spectral graph clustering via convex layer aggregation: Theory and algorithms. *IEEE Transactions on Signal and Information Processing over Networks*. 2017;**3**:553-567

[39] Lee KC, Ho J, Kriegman DJ. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005;**27**:684-698

[40] Liu G, Lin Z, Yan S, Sun J, Yu Y, Ma Y. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2012;**35**: 171-184

[41] Hu S, Short NJ, Riggan BS, Gordon C, Gurton KP, Thielke M, et al. A polarimetric thermal database for face recognition research. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Las Vegas, NV, USA: IEEE Conference on Computer Vision and Pattern Recognition, CVPR; 2016. pp. 187-194

[42] Hunter B, Strohmer T. Performance analysis of spectral clustering on compressed, incomplete and inaccurate measurements. 2010. arXiv preprint arXiv:1011.0997