Deep Self-Supervised t-SNE for Multi-modal Subspace Clustering

Qianqian Wang
Xidian University
Xi'an, Shaanxi, China
qqwang@xidian.edu.cn

Wei Xia
Xidian University
Xi'an, Shaanxi, China
xd.weixia@gmail.com

Zhiqiang Tao
Santa Clara University
Santa Clara, CA, USA
ztao@scu.edu

Quanxue Gao*
Xidian University
Xi'an, Shaanxi, China
qxgao@xidian.edu.cn

Xiaochun Cao
Chinese Academy of Sciences
Beijing, China
caoxiaochun@iie.ac.cn

ABSTRACT

Existing multi-modal subspace clustering methods, aiming to exploit the correlation information between different modalities, have achieved promising preliminary results. However, these methods might be incapable of handling real problems with complex heterogeneous structures between different modalities, since the large heterogeneous structure makes it difficult to directly learn a discriminative shared self-representation for multi-modal clustering. To tackle this problem, in this paper, we propose a deep Self-supervised t-SNE method (StSNE) for multi-modal subspace clustering, which learns soft label features by multi-modal encoders and utilizes the common label feature to supervise soft label feature of each modal by adversarial training and reconstruction networks. Specifically, the proposed StSNE consists of four components: 1) multi-modal convolutional encoders; 2) a self-supervised t-SNE module; 3) a self-expressive layer; 4) multi-modal convolutional decoders. Multi-modal data are fed to encoders to obtain soft label features, for which the self-supervised t-SNE module is added to make full use of the label information among different modalities. Simultaneously, the latent representations given by encoders are constrained by a self-expressive layer to capture the hierarchical information of each modal, followed by decoders reconstructing the encoded features to preserve the structure of the original data. Experimental results on several public datasets demonstrate the superior clustering performance of the proposed method over state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Cluster analysis**; **Learning latent representations**.

KEYWORDS

multi-view clustering, subspace learning, adversarial learning

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475319

ACM Reference Format:
Qianqian Wang, Wei Xia, Zhiqiang Tao, Quanxue Gao*, and Xiaochun Cao. 2021. Deep Self-Supervised t-SNE for Multi-modal Subspace Clustering. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3474085.3475319

1 INTRODUCTION

Clustering analysis is a fundamental task in a wide range of fields, such as machine learning, pattern recognition, computer vision, and data mining [19]. There are numerous works proposed on this topic, among which, multi-modal clustering [35] is of particular interest due to the ubiquitous multi-modal data existing in real-world applications. Multi-modal data describe the objects' characteristics from distinct perspectives. For example, an image could be characterized by various descriptors, such as color, depth, structure, etc. In the past few years, multi-modal clustering (MMC) methods [6, 12] have been developed rapidly by exploring the complementary information among multiple modalities. Existing MMC methods can be roughly divided into two categories: traditional methods and deep methods. The traditional methods such as the non-negative matrix factorization (NMF) based methods [14, 20], multi-kernel learning (MKL) [15, 37] methods, subspace methods [7, 23], and the graph-based methods [21, 22, 29] mainly adopt shallow and linear embedding functions to reveal the intrinsic structure of data. However, they are difficult to depict the structure of high-dimensional nonlinear data. In addition, they may suffer from the curse of dimensionality.

To address these problems, deep learning methods are developed to deal with the multi-modal clustering problem. For example, Andrew et al. [2] proposed a deep canonical correlation analysis (DCCA) method to learn complex nonlinear transformations of two-modal data such that the resulting representations are highly linearly correlated. Abavisani and Patel [1] employed convolutional neural networks for unsupervised multi-modal subspace clustering (DMSC). Although these methods have achieved promising results, they are still limited by the complex heterogeneous information between different modalities. To name a few, two challenges could be raised by multi-modal subspace clustering:

- How to consider the distribution of the inter-modal data and the intra-modal data simultaneously to learn a representative shared subspace to improve clustering accuracy?

- How to guarantee that the features extracted from multiple modalities contain more discriminative information and benefit the performance of the clustering task?

To address the aforementioned challenges, as shown in Figure 1, we propose a novel Self-supervised t-SNE method (StSNE) for multi-modal subspace clustering to improve multi-modal clustering performance. The proposed method consists of four parts: multi-modal convolutional encoders, a self-supervised t-SNE module, a self-expressive layer, and multi-modal convolutional decoders. The multi-modal convolutional encoders map each modality's high-dimensional data to a low-dimensional subspace to obtain latent representations, and meanwhile, the self-supervised t-SNE module constrains soft label features with a boosted consensus cluster distribution further. On the derived soft labels, the self-expressive layer is employed to learn a common soft label feature shared by all modalities. Thus, the intrinsic information from intra-modal and inter-modal can be well revealed. Finally, we utilize multi-modal convolutional decoders to reconstruct the original data, which ensures the representation produced by the encoder can well reflect the characteristics of the original data. The main contributions of our method are summarized as:

- We propose a novel self-supervised t-SNE method (StSNE) for multi-modal subspace clustering by developing a self-expression layer to consider both inter-modal and intra-modal distributions.
- The self-supervised t-SNE module makes soft label features of each modality similar to each other, which helps achieve a more compact clustering structure and thus improves the clustering performance.
- A deep convolutional encoder-decoder architecture is designed and developed to implement multi-modal data reconstruction, where the encoded features enable us to capture the overall structure distribution of original data.
- We conduct extensive experiments and comparisons with state-of-the-art works to demonstrate the effectiveness of the proposed method.

2 RELATED WORK

Multi-modal clustering aims to extract a common representation for multiple modalities. One representative method is based on canonical correlation analysis (CCA) [2, 4, 26]. It makes two-modal data similar to each other by maximizing the correlation of subspaces of two modalities, which can learn a consistent representation for multi-modal data. Following CCA, Andrew et al. [2] proposed a DNN extension based on CCA (DCCA) to learn complex nonlinear transformations of two-modal data such that the resulting representations are highly linearly correlated. Inspired by CCA and reconstruction-based objectives, the deep canonically correlated auto-encoder (DCCAE) was developed in [26]. Different from DCCA, DCCAE optimizes the combination of canonical correlation between the learned "bottleneck" representation and the reconstruction errors of the auto-encoders. However, all these methods are limited to two-modal data. To overcome this challenge, Benton et al. [4] presented Deep Generalized Canonical Correlation Analysis (DGCCA), a method to learn nonlinear transformations of arbitrarily multiple modalities of data, such that the resulting transformations are maximally informative of each other.

Apart from the CCA-based methods, a variety of multi-modal embedding clustering methods have been proposed based on auto-encoders [13, 24, 25]. For example, Abavisani and Patel [1] employed convolutional neural networks for unsupervised multi-modal subspace clustering (DMSC). Though this method achieves promising results, it is not straightforward to apply DMSC on large datasets due to the self-expression constraint. For another example, Lin et al. [13] proposed a novel joint framework for deep multi-modal clustering (DMJC), where multiple deep embedded features, multi-modal fusion mechanism, and clustering assignments are learned simultaneously. Li et al. [12] developed deep multi-view clustering via generative adversarial networks (GAN).

Though the aforementioned methods have achieved promising results, they ignore the distribution consistency of the inter-modal data. In addition, due to the complex heterogeneous information between different modalities, the shared representation directly recovered from each modal is difficult to explore the intrinsic clustering feature.

3 SELF-SUPERVISED T-SNE FOR MULTI-MODAL SUBSPACE CLUSTERING

3.1 Motivations

To address the limitation of existing multi-modal subspace clustering, we focus on the distribution consistency of the inter-modal data. If the distribution of each modal is closed to an "ideal" consensus distribution that has a better clustering structure, the intrinsic clustering feature can be well explored. It is well known that t-distributed stochastic neighbor embedding (t-SNE) [17] is an effective technique to encourage a compact data structure in the low-dimensional embedding space. Motivated by it, we utilize t-SNE to model the consensus distribution across different modalities and push each modal towards the consensus one. As a result, the correlation of different modalities can be well exploited.

Additionally, the structures embedded in each modal are also important for clustering. To this end, we utilize self-representation learning to recover a common representation shared by all the modalities. Considering the possible noise and high-dimensionality of the original multi-modal data, we utilize convolutional encoders to extract soft label features of each modal. Correspondingly, convolutional decoders are utilized to make the obtained latent representation well preserve the intrinsic structure of original data.

3.2 The Framework of StSNE

The framework of the proposed StSNE is shown in Fig. 1, which consists of multi-modal convolutional encoders, a self-supervised t-SNE module, a consistent self-expressive layer, and multi-modal convolutional decoders. Our model aims to partition a set of N data points into k clusters by using multi-modal features $\{\mathbf{X}^{(m)}\}$, where $\mathbf{X}^{(m)} = \{x_1^{(m)}, \ldots, x_N^{(m)}\}$ denotes the original features of the m-th modality, $1 \leq m \leq M$, and M represents the number of modalities. In each $\mathbf{X}^{(m)}$, $x_i^{(m)}$ denotes the i-th sample, $1 \leq i \leq N$. We use d_m to denote the feature dimension of the samples in the m-th modality.

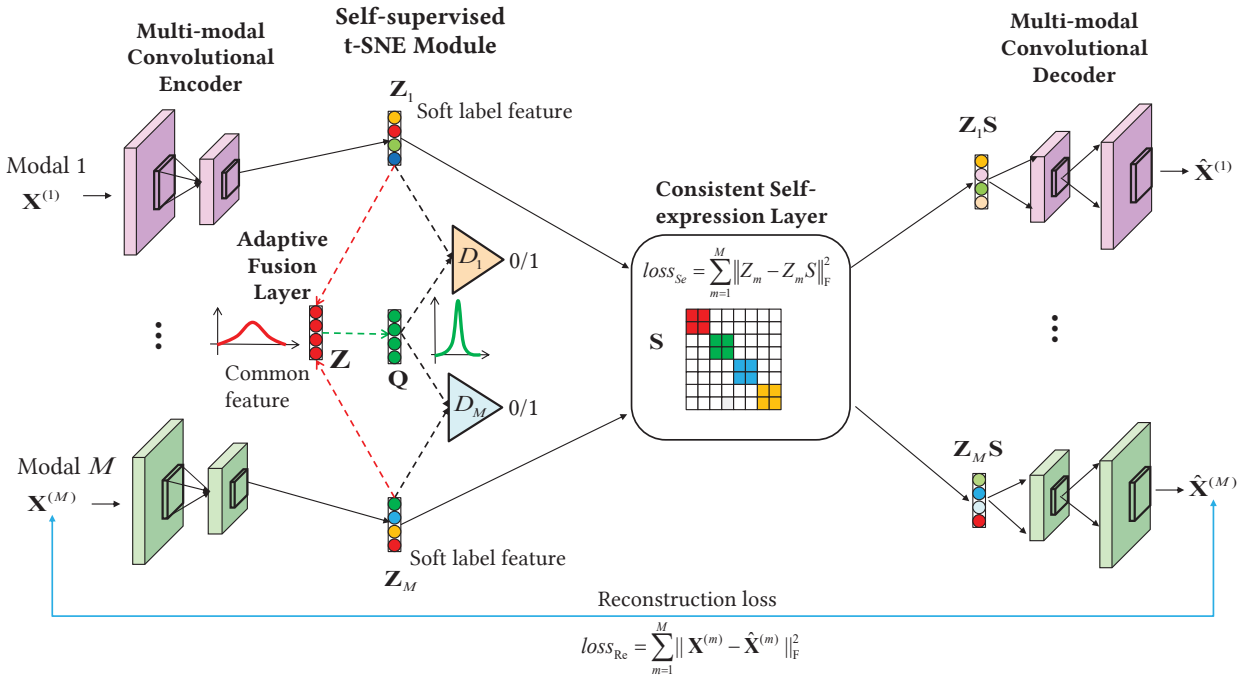


Figure 1: Network structure diagram of the our Method, which consists of multi-modal convolutional encoders, a self-supervised t-SNE module, a consistent self-expressive layer and multi-modal convolutional decoders.

3.2.1 Multi-modal Convolutional Encoders. For the m-th modal data $\mathbf{X}^{(m)}$, the multi-modal convolutional encoder aims to learn a non-linear mapping $h(\mathbf{X}^{(m)}; \theta_m)$ which can transform the original features to a soft label feature $\mathbf{Z}_m = \{z_1^{(m)}, \ldots, z_N^{(m)}\} (\mathbf{Z}_m \in \mathbb{R}^{k \times N})$, where k is the output dimension of the convolutional encoder, which is also the number of clusters. Specifically, it maps the d_m-dimensional input data $x_i^{(m)}$ to a k-dimensional representation $z_i^{(m)}$. This mapping could be obtained by $h(\mathbf{X}^{(m)}, \theta_m) = \mathbf{Z}_m$, where $h(\cdot)$ refers to the encoder mapping function parameterized by θ_m.

3.2.2 Self-supervised t-SNE Module. After multi-modal convolutional encoders, we derive M soft label features $\mathbf{Z}_m$. In order to make each modal's soft label feature as close as possible, we construct a common soft label feature and make each modal's soft label feature more similar to the common soft label feature distribution. Therefore, we obtain the common soft label feature $\mathbf{Z}$ by $\mathbf{Z} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{Z}_m$.

We obtain the common label soft feature distribution $\mathbf{Q}$ with the t-student distribution of t-SNE [17] between the common soft label feature $\mathbf{Z}$ and the common cluster centroids $\{\mu_j\}_{j=1}^{k}$. In our model, we use mean square error and adversarial training to make each modal's soft label feature close to the common representation distribution. This allows the latent representation of each modality to be closer to the same and to make the latent layer features belonging to the same clusters more similar.

Based on the above analysis, the common cluster centroids $\{\mu_j\}_{j=1}^{k}$ and each modal discriminator $\mathbf{D}_m$ constitute a self-supervised t-SNE module. Each discriminator $\mathbf{D}_m$ consists of 3 fully connected layers, and $z_i^{(m)} \sim P(\mathbf{Z}_m)$ is a generated sample and $q_i \sim P(\mathbf{Q})$ is a real instance, where the notation $x \sim P(\mathbf{X})$ represent x is a sample of $\mathbf{X}$. $\mathbf{D}_m$ feeds back the result to the generator network and updates the parameters of the generator. By this means, the discriminator works as a regularizer to guide the training of our multi-modal encoders, which enhances the robustness of embedding representation and avoids the over-fitting issue effectively.

3.2.3 Consistent Self-expressive Layer. Some self-expressiveness based methods [1, 9] have attracted much attention, which aims to express the data point as a linear combination of other points in the same subspace. We obtain the soft label feature $\mathbf{Z}_m$ from multi-modal convolutional encoders and send them to the self-expression layer. In the same space, one data point can be represented linearly by other data points. Then we can get the $\mathbf{Z}_m\mathbf{S} = \mathbf{Z}_m$, where $\mathbf{S}$ is the self-representation coefficient matrix. M modalities share a same self-expression coefficient matrix. In order to prevent the trivial solution $\mathbf{S} = \mathbf{I}$, we constraint $diag(\mathbf{S}) = 0$. Then we can leverage the matrix $\mathbf{S}$ to construct the affinity matrix by the following equation $\mathbf{C} = \frac{1}{2}(|\mathbf{S}| + |\mathbf{S}^T|)$, Finally, we apply $\mathbf{C}$ for spectral clustering [18].

3.2.4 Multi-modal Convolutional Decoders. To learn a better soft label feature $\mathbf{Z}_m$, we add multi-modal convolutional decoders. It has an opposite architecture to the multi-modal convolutional encoder and could reconstruct the m-th modal data from the soft label feature $\mathbf{Z}_m$. Denote $\hat{\mathbf{X}}^{(m)} = g(\mathbf{Z}_m\mathbf{S}, \delta_m)$, where $\hat{\mathbf{X}}^{(m)}$ represents the reconstructed sample matrix of the m-th modal, $g(\cdot)$ refers to decoder mapping function parameterized by δ_m.

3.3 Loss Function Analysis

The total loss function of our model is defined as follows:

$$Loss = \min_{\theta,\mu,\mathrm{S},\delta} \max_{\mathbf{D}_m} L_{AT} + \lambda_1 L_{Se} + \lambda_2 L_{Re}, \tag{1}$$

which is composed of three parts: the Self-supervised t-SNE loss L_{AT}, the Self-expression loss L_{Se}, and the Reconstruction loss L_{Re}. λ_1 and λ_2 are two parameters to balance the impact of the Self-expression loss and the Reconstruction loss. θ are the encoder parameters, δ are the decoder parameters, $\mathbf{S}$ is the self-representation coefficient matrix, μ are the common cluster centroids and $\mathbf{D}_m$ is m-th discriminator. The common clustering centroids μ are initialized by K-means on the common soft label feature $\mathbf{Z}$ and updated by backpropagating gradients along with the network training.

3.3.1 Self-supervised t-SNE Loss. The t-distributed stochastic neighbor embedding (t-SNE) [17] is a nonlinear dimensionality reduction algorithm for exploring high-dimensional data. It maps multidimensional data to two or more dimensions suitable for human observation. The main idea is to use conditional probability to represent the similarity of the distances of the high-dimensional distribution points, and the points of the low-dimensional distribution are also represented. As long as the conditional probabilities of the two are very close, it means that the points of the high-dimensional distribution have been mapped to the low-dimensional distribution. We will draw on the idea of t-SNE to constrain the distribution of each modal soft label feature to capture the similarity of the data between modals.

The self-supervised t-SNE loss function is defined as follows:

$$L_{AT} = \min_{\theta,\mu} \max_{\mathbf{D}_m} L_T + \lambda_3 L_A, \tag{2}$$

consisting of the t-SNE loss L_T and the adversarial loss L_A. λ_3 is a balancing parameter between L_T and L_A.

t-SNE loss: Given the initial cluster centroids $\{\mu_j\}_{j=1}^k$, we use the Student's t-distribution in t-SNE [17] as a kernel to measure the similarity between common latent representation point z_i and centroid μ_j:

$$q_{ij} = \frac{(1 + ||z_i - \mu_j||^2)^{-1}}{\sum_{j'} (1 + ||z_i - \mu_{j'}||^2)^{-1}}, \tag{3}$$

where q_{ij} is interpreted as the probability of assigning sample i to cluster j, i.e., soft assignment. In our model, we use mean square error and adversarial training to make each modal soft label feature close to the common soft label feature distribution. Hence, our t-SNE loss function is given by

$$L_T = \min_{\theta,\mu} \sum_{m=1}^{M} ||\mathbf{Z}_m - \mathbf{Q}||_F^2, \tag{4}$$

where $||\cdot||_F$ denotes the matrix Frobenius norm. We constrain the soft label feature $\mathbf{Z}_m$ to the common soft label feature distribution $\mathbf{Q}$ by the t-SNE loss L_T, and, ideally, each view should exhibit the same cluster structure. However, since individually using the MSE implementation of L_T leverages one dimensional, element-wise error model and ignores the relationship between them. Hence, we further use an adversarial loss [12] to model the cluster distributions.

Adversarial loss: The adversarial loss function is given by

$$\begin{aligned} L_A = \min_{\theta,\mu} \max_{\mathbf{D}_m} \sum_{m=1}^{M} (&\mathbb{E}_{q\sim P(\mathbf{Q})}[log\mathbf{D}_m(q)] \\ &+\mathbb{E}_{z^m\sim P(\mathbf{Z}_m)}[log(1-\mathbf{D}_m(z^m))]) \end{aligned}, \tag{5}$$

where the notation $\mathbb{E}$ represents $\mathbb{E}_{x\sim P(X)}[f(x)] = \frac{1}{N}\sum_{i=1}^{N} f(x^i)$, and N denotes the number of samples. We treat $P(\mathbf{Q})$ as "real" samples since $\mathbf{Q}$ enjoys a sharper and more consensus clustering distribution than each single view. Thus, we could adopt adversarial training to push each view's soft labels $\mathbf{Z}_m$ towards view-consensus.

We update the multi-modal convolutional encoders θ and the common cluster centroids μ, as well as the M discriminator networks $\mathbf{D_m}$ by optimizing the self-supervised t-SNE loss L_{AT}.

3.3.2 Self-expression Loss. In the self-expression layer, to better perform the self-expression property and acquire a better self-expression coefficient matrix $\mathbf{S}$, we minimize the self-expression loss function:

$$\begin{aligned} L_{Se} = \min_{\theta,\mathrm{S}} ||\mathrm{S}||_1 + \sum_{m=1}^{M} ||\mathbf{Z}_m - \mathbf{Z}_m\mathrm{S}||_F^2, \\ s.t., diag(\mathrm{S}) = 0. \end{aligned} \tag{6}$$

where $||\cdot||_1$ denotes the matrix L_1 norm, and $||\mathrm{S}||_1$ is the regular loss function for the coefficient matrix $\mathbf{S}$. We update the multi-modal convolutional encoder and the consistent self-expressive layer by minimizing the self-expression loss.

3.3.3 Reconstruction Loss. In order to guarantee the effectiveness of the representation processed by the multi-modal convolutional encoder and the self-expression layer, we add the multi-modal convolutional decoder to reconstruct data. The soft label feature $\mathbf{Z}_m\mathbf{S}$ from the self-expression layer are fed to the multi-modal convolutional decoder and we can acquire the reconstruct data. Minimize errors between reconstructed and original data to optimize the network. Therefore, the reconstruction loss for the network is

$$L_{\mathrm{Re}} = \min_{\theta,\mathrm{S},\delta} \sum_{m=1}^{M} ||\mathbf{X}^{(m)} - \hat{\mathbf{X}}^{(m)}||_F^2. \tag{7}$$

The reconstruction loss guarantees the reliability of the latent representation by updating the multi-modal convolutional encoder, the consistent self-expressive layer, and the multi-modal convolutional decoder.

3.4 Model Training

In the proposed model, we use two steps to train the model and optimize the network parameters.

Step. 1: We pre-train the network using Eq (7). We send the multi-modal data $\mathbf{X}^{(m)}$ to the multi-modal convolutional encoder and obtain the reconstruction data from the multi-modal convolutional decoder. In the pre-training step, we set the learning-rate to 0.001 and minimize the error between the original data and the reconstruction data to optimize the network and update encoder parameter θ and decoder parameter δ, where we use mean squared error (MSE) [27] to optimize the objective function.

Step. 2: We train the entire network using Eq (1). We use the multi-modal convolutional encoder parameter θ and the multi-modal convolutional decoder parameter δ from the first step training to train the entire network, i.e., minimizing the total loss to

Table 1: Statistics of four multi-modal real-world datasets. Note that the training and testing images in each dataset are jointly utilized for clustering.

Dataset	#Sample	#Class	#Modality	#Size
Fashion MNIST	70000	10	2	$28 \times 28 \times 1$
COIL20	1440	20	2	$128 \times 128 \times 1$
YTF	10000	41	3	$55 \times 55 \times 3$
FRGC	2462	20	3	$32 \times 32 \times 3$

update model parameters θ, δ, μ and the coefficient matrix $\mathbf{S}$. We obtain the shared coefficient matrix $\mathbf{S}$ from self-expression layer, and calculate the affinity matrix $\mathbf{C} = \frac{1}{2}(|\mathbf{S}| + |\mathbf{S}|^{\mathrm{T}})$. Finally, we use the affinity matrix $\mathbf{C}$ and spectral clustering method to complete data clustering.

4 EXPERIMENTS

In this section, we first give the basic settings in our experiments and then provide the clustering performance on four real-world datasets as well as some experimental analyses.

4.1 Experimental Settings

4.1.1 Datasets. We construct four new multi-modal datasets by generating auxiliary modalities upon the original images, each of which is described as follows.

- **Fashion-MNIST dataset**: Fashion-MNIST [30] is a widely-used benchmark dataset consisting of 70, 000 fashion product images with 28×28 pixels. In our experiment, we use the original image features as the first mode and the edge features of the extracted fashion product as the second mode.
- **COIL20 dataset**: COIL20 [11] collects 1440 128×128 grayscale object images of 20 categories viewed from varying angles. Like the Fashion-MNIST dataset, we use the original picture feature as the first mode and the edge feature of the extracted object as the second mode.
- **Youtube-Face (YTF) dataset**: Following [34], we choose the first 41 subjects of YTF dataset. Faces inside images are first cropped and then resized to 55 by 55 sizes [28]. In this paper, we implement its original RGB picture as the first mode, the gray picture converted from the original RGB picture as the second mode, and the extracted edge feature as the third mode.
- **FRGC dataset**: Using 20 random selected subjects in [34] from the original dataset, we collect 2,462 face images. Similarly, we first crop the face regions and resize them into 32 by 32 images. We treat the dataset in the same way as the YTF dataset.

The examples and experimental statistics of four multi-modal datasets are shown in Figure 2 and Table 1. Notably, both the training and testing images in each dataset are utilized for unsupervised clustering.

4.1.2 Comparison Algorithms. We choose two single-modal clustering methods: **K-means** clustering [8] and Deep Embedding Clustering (**DEC**) [32]; four traditional multi-view clustering methods: Robust Multi-View K-Means Clustering (**RMKMC**) [5], Binary Multi-View Clustering (**BMVC**) [36], Multiview clustering by joint latent representation and similarity learning(**LALMVC**) [31], and Consistent and specific multi-view subspace clustering(**CSMSC**) [16]; two deep two-modal clustering methods: Deep Canonical Correlation Analysis (**DCCA**) [2] and Deep Canonically Correlated Auto-Encoders (**DCCAE**) [26]; three deep multi-modal clustering methods: Deep Generalized Canonical Correlation Analysis (**DGCCA**) [4], Joint framework for Deep Multi-view Clustering (**DMJC**) [13] (we adopt the second scheme DMJC-T as the baseline), and Deep Multimodal Subspace Clustering (**DMSC**) [1].

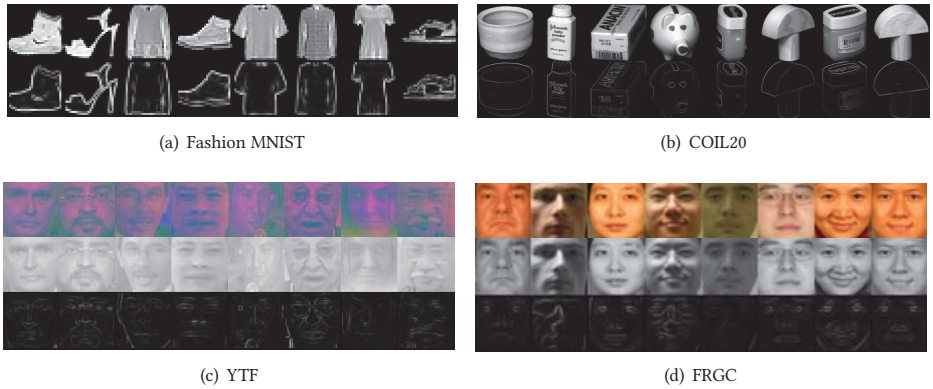
4.1.3 Implementation Details. We implement our method and other non-linear methods with the public toolbox of PyTorch. We run all the experiments on the platform of Ubuntu Linux 16.04 with NVIDIA Titan Xp Graphics Processing Units (GPUs) and 64 GB memory size. We select Adam [3] optimizer with default parameter setting to train our model and fix the learning rate at 0.001. We conduct 2000 epochs for the first step train and we conduct 3000 epochs for the second step train. Our method and DMSC have a limitation of full-batch training since it utilizes a self-expressive layer as regularization. Therefore, it cannot handle large-scale data efficiently. In our experiments, because Fashion MNIST dataset and YTF dataset are large (all larger than 10,000), we randomly select 2000 samples on these two datasets for experiments. The batch size is set as the number of sample size. All the other linear methods are tested under the same environment by Matlab.

Since DCCA and DCCAE can only deal with two modalities, we choose the best two modalities in our models according to their performance as the two branches for DCCA and DCCAE. After multi-modal feature learning, we concatenate the embedding features in two branches to perform K-means. For DGCCA, we use the shared representation to perform K-means directly. The pre-trained network parameters of DCCA, DCCAE, and DGCCA are also kept consistent with our models. For the DMJC algorithm, we choose the same multi-modal convolutional encoder as our model as the multi-modal branch of it.

4.1.4 Modal Network. When designing the network, for each convolutional auto-encoder, we set up a three-layer network. The size of the first-layer convolution kernel is $4 \times 4 \times 10$, and the step size is 2. The size of the second-layer convolution kernel is $3 \times 3 \times 20$, and the step size is 1. The size of the third-layer convolution kernel is $4 \times 4 \times 30$, and the step size is 2. The deconvolution decoder has a convolution kernel size opposite to that of the convolution auto-encoder. Specifically, for the YTF dataset, the second layer output is zero-padded to match the dimension.

4.2 Experimental Results

4.2.1 Comparison with Baselines. In order to evaluate the performance of clustering algorithms, we adopt two metrics, *i.e.,* clustering accuracy (ACC) [10] and normalized mutual information (NMI) [33], to measure its clustering performance by comparing it



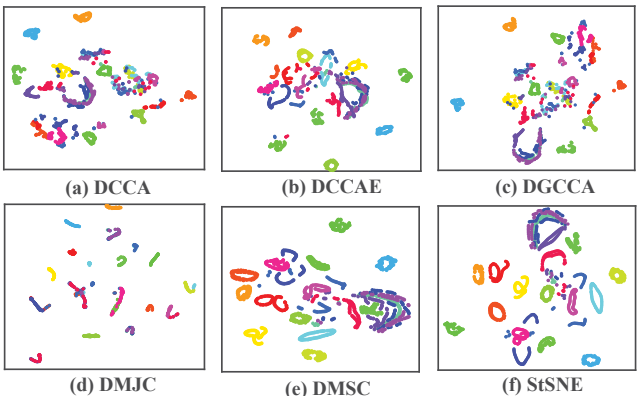
(a) Fashion MNIST (b) COIL20

(c) YTF (d) FRGC

Figure 2: The examples of four multi-modal datasets.

Table 2: The optimal clustering accuracy(ACC %) and the normalized mutual information (NMI %) on all the datasets. Best results are highlighted in bold.

Methods	Fashion-MNIST		COIL-20		FRGC		YTF	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
K-means [8]	51.27	49.99	57.49	73.22	23.62	27.12	56.01	75.23
DEC [32]	51.80	54.60	68.00	80.25	37.80	50.50	37.10	44.60
RMKMC [5]	53.32	52.87	60.97	74.93	23.52	25.85	57.21	74.56
BMVC [36]	45.36	38.05	34.31	40.33	41.51	45.92	28.13	38.28
LALMVC [31]	57.20	59.31	64.79	76.83	49.68	57.27	40.85	49.6
CSMSC [16]	62.45	61.80	62.08	73.15	52.15	66.23	54.65	73.55
DCCA [2]	52.74	53.82	55.76	64.91	22.91	24.75	45.19	60.35
DCCAE [26]	51.87	53.01	61.60	71.56	32.33	31.22	45.57	60.15
DGCCA [4]	56.28	57.04	54.01	62.40	23.76	24.53	47.26	61.38
DMJC [13]	61.41	63.41	72.99	81.58	44.07	59.79	61.15	77.40
DMSC [1]	59.55	65.07	74.10	86.82	72.83	80.96	62.80	80.16
StSNE	**65.65**	**68.99**	**82.43**	**91.72**	**74.33**	**81.44**	**67.20**	**83.44**



(a) DCCA (b) DCCAE (c) DGCCA

(d) DMJC (e) DMSC (f) StSNE

Figure 3: Visualization of the common latent representation given by different methods with t-SNE on the COIL20 dataset.

with nine baseline methods on four datasets. The higher ACC/NMI values indicate better clustering results.

Table 2 reports the clustering performances of all the compared methods on four datasets. From the comparison results, we have the following observations.

1) The proposed StSNE model consistently achieves the best performance on four datasets in terms of both ACC and NMI, which clearly supports the improved clustering performance brought by the self-supervised t-SNE module among inter-modal and the self-expressive layer of intra-modal data.

2) Our proposed model significantly outperforms the single-modal clustering methods among most cases. For example, on the FRGC dataset, the performances of K-means and DEC are only 23.62% and 37.80% by ACC, and 50.50% by NMI. This is because the single-modal method does not consider the information of other modalities and the clustering structure they learn cannot fully reveal the data characteristics.

3) The ACC and NMI of DCCA are only 22.91% and 24.75% on the FRGC dataset, which is the worst result. It is probably because that DCCA cannot ensure the representation after the encoder network still reflects the distribution of the original data. DCCAE also obtains poor ACC and NMI results of 32.33% and 31.22% on

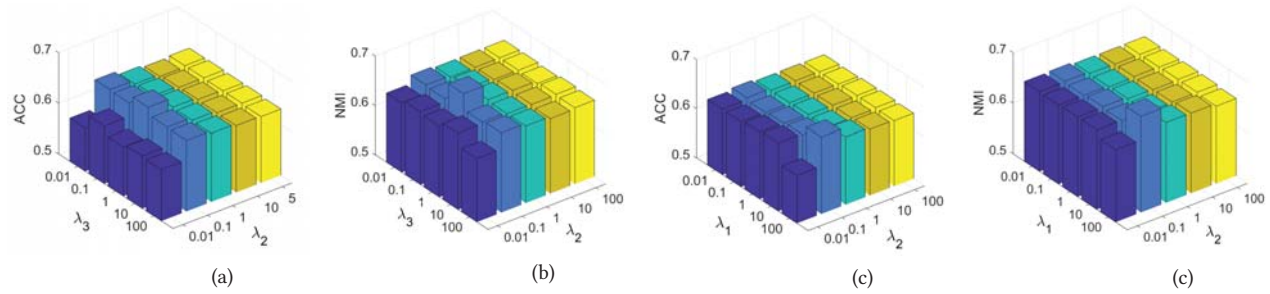


Figure 4: Influence of parameter changes on clustering performance on Fashion-MNIST dataset.

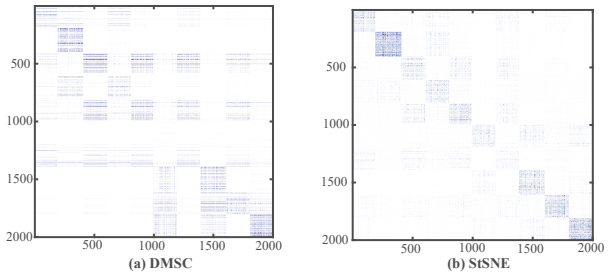


Figure 5: Visualization of the learned subspace representation for a) DMSC and b) our Method on Fashion-MNIST dataset.

Table 3: Ablation Study on Fashion-MNIST dataset in terms of ACC (%) and NMI (%). Best results are highlighted in bold.

Methods	ACC	NMI
StSNE w/o L_{Re}	52.35	59.74
StSNE w/o L_{Se}	63.50	65.93
StSNE w/o L_{AT}	64.15	65.68
StSNE (full model)	**65.65**	**68.99**

the FRGC dataset, which may be because it doesn't consider the relationships among intra-modal data.

4) Even though DMSC which is designed for multi-modal data clustering performs better than other methods in most cases, our proposed model also has an improvement over DMSC. It is because that DMSC may not make full use of the information among the inter-modal data. To sum up, the proposed StSNE has shown a promising clustering performance compared with state-of-the-art multi-modal clustering methods, since our approach not only constrains the intra-modal data similarity distribution through the self-expression layer but also constrains the inter-modal data similarity distribution by the self-supervised t-SNE module.

4.2.2 Ablation Study. In this subsection, we perform a detailed ablation study on our model regarding different loss functions. The overall objective function of StSNE consists of three parts: the reconstruction loss L_{Re}, the self-expression loss L_{Se}, and self-supervised t-SNE loss L_{AT}. On the Fashion-MNIST dataset, we remove each loss function in turn and conduct the experiment. As shown in Table 3, we can find that: 1) the self-supervised t-SNE module has a certain impact on clustering performance, which constrains the inter-modal data distribution more closely and maximizes the cluster representation of each modal's latent features to obtain a better common subspace representation; 2) the self-expressive layer has a significant effect on the proposed model, and the correlation among intra-modal data play an important role in the improvement of clustering performance; 3) the deep convolutional encoder-decoder has the biggest impact on the proposed method, whose role is to ensure the overall structure of the data and make the encoded data reliable. The above observations indicate that all the three components in our proposed StSNE model are designed reasonably.

4.2.3 Visualization. Figure 3 provides the t-SNE [17] visualization of feature embeddings obtained by five competitive compared methods and our proposed method on the COIL20 dataset. In detail, we apply t-SNE on the common-modal feature representations given by different methods, respectively. As can be seen, our approach exhibits a more clear and compact cluster structure than all the other methods. As shown in Figure 5, in order to illustrate the role of each module more clearly, we visualize the learned subspace representation for each method. Figure 5(a) displays the visualization result of DMSC, in which there are a lot of noise points resulting in a low clustering performance. For our method, its clustering performance significantly improves as shown in Figure 5(b) due to that it can capture the correlation between multi-modal data to learn a more consistent subspace representation.

4.2.4 Parameters Analysis. In our model, there are three regularization parameters λ_1, λ_2 and λ_3. We use the method of controlling variables to analyze the parameters. Firstly, we fix the regularization parameter λ_3 of the adversarial loss and vary both the regularization parameters λ_1 and $\lambda_2$of the self-expression loss and the reconstruction loss in the range of $\{0.01, 0.1, 1, 10, 100\}$. Then, we fix λ_1 and vary λ_2 and λ_3 in the range of $\{0.01, 0.1, 1, 10, 100\}$. Since the strategies of setting parameters are the same on all the four datasets, we only show the effect of parameters on Fashion-MNIST dataset for simplicity. From Figure 4, we can notice that 1) our method can achieve the best ACC and NMI results on Fashion-MNIST dataset when $\lambda_1 = 100, \lambda_2 = 0.1$ and $\lambda_3 = 1$; 2) our method is stable since varying parameters has little influence on the clustering performance.

5 CONCLUSIONS

In this paper, we proposed a novel multi-modal clustering method, namely Self-supervised t-SNE (StSNE) for multi-modal subspace clustering. The proposed model derives a soft label feature of each modal by a convolutional encoder, utilizes a self-supervised t-SNE module to make the distribution of the learned soft label feature close to the ideal distribution of data, employs a self-expressive layer to recover a shared representation, and simultaneously performs data reconstruction via a convolutional decoder. Thus, the correlation distribution information of both the intra-modal data and the inter-modal data is effectively captured. Consequently, the recovered soft label feature that is shared by all modalities can well reveal the intrinsic structure of multi-modal data. Experimental results on four real-world multi-modal image datasets demonstrate the superiority of our model over several state-of-the-art multi-modal clustering methods.

6 ACKNOWLEDGMENTS

This work is supported by the the National Key R&D Program of China under Grant 2020AAA0109304, National Natural Science Foundation of China under Grant 61773302, Initiative Postdocs Supporting Program BX20190262, China Postdoctoral Science Foundation (Grant 2019M663642), the National Natural Science Foundation of Shaanxi Province (Grant 2020JZ-19, 2020JQ-327).

REFERENCES

[1] Mahdi Abavisani and Vishal M Patel. 2018. Deep multimodal subspace clustering networks. *IEEE Journal of Selected Topics in Signal Processing* 12, 6 (2018), 1601–1614.
[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *ICML*. 1247–1255.
[3] Diederik P arXiv preprint Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[4] Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, and Raman Arora. 2017. Deep generalized canonical correlation analysis. *arXiv preprint arXiv:1702.02519* (2017).
[5] Xiao Cai, Feiping Nie, and Heng Huang. 2013. Multi-view k-means clustering on big data. In *IJCAI*.
[6] Jiafeng Cheng, Qianqian Wang, Zhiqiang Tao, and Quanxue Gao. 2020. Multi-View Attribute Graph Convolution Networks for Clustering. In *IJCAI*. 2973–2979.
[7] Quanxue Gao, Huanhuan Lian, Qianqian Wang, and Gan Sun. 2020. Cross-Modal Subspace Clustering via Deep Canonical Correlation Analysis.. In *AAAI*. 3938–3945.
[8] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108.
[9] Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. 2017. Deep subspace clustering networks. In *NIPS*. 24–33.
[10] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2, 1-2 (1955), 83–97.
[11] Fengfu Li, Hong Qiao, and Bo Zhang. 2018. Discriminatively boosted image clustering with fully convolutional auto-encoders. *PR* 83 (2018), 161–173.
[12] Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, and Zhaohua Yang. 2019. Deep adversarial multi-view clustering network. In *IJCAI*. 2952–2958.
[13] Bingqian Lin, Yuan Xie, Yanyun Qu, and Cuihua Li. 2018. Deep multi-view clustering via multiple embedding. *CoRR* (2018).
[14] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. 2013. Multi-view clustering via joint nonnegative matrix factorization. In *ICDM*. 252–260.
[15] Xinwang Liu, Yong Dou, Jianping Yin, Lei Wang, and En Zhu. 2016. Multiple kernel k-means clustering with matrix-induced regularization. In *AAAI*. 1888ÍC1894.
[16] Shirui Luo, Changqing Zhang, Wei Zhang, and Xiaochun Cao. 2018. Consistent and specific multi-view subspace clustering. In *AAAI*, Vol. 32.
[17] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 2579–2605.
[18] Andrew Y Ng, Michael I Jordan, and Yair Weiss. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*. 849–856.
[19] Gan Sun, Yang Cong, Jiahua Dong, Yuyang Liu, Zhengming Ding, and Haibin Yu. 2021. What and How: Generalized Lifelong Spectral Clustering via Dual Memory. *IEEE TPAMI* (2021).
[20] Gan Sun, Yang Cong, Yulun Zhang, Guoshuai Zhao, and Yun Fu. 2020. Continual multiview task learning via deep matrix factorization. *IEEE TNNLS* 32, 1 (2020), 139–150.
[21] Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. 2017. From ensemble clustering to multi-view clustering. In *IJCAI*.
[22] Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. 2020. Marginalized Multiview Ensemble Clustering. *IEEE TNNLS* 31, 2 (2020), 600–611.
[23] Qianqian Wang, Jiafeng Cheng, Quanxue Gao, Guoshuai Zhao, and Licheng Jiao. 2020. Deep Multi-view Subspace Clustering with Unified and Discriminative Learning. *IEEE TMM* 99, 9 (2020), 1–11.
[24] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. 2018. Partial multi-view clustering via consistent GAN. In *IEEE ICDM*. 1290–1295.
[25] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. 2021. Generative Partial Multi-View Clustering With Adaptive Fusion and Cycle Consistency. *IEEE TIP* 30 (2021), 1771–1783.
[26] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2016. On deep multi-view representation learning: objectives and optimization. *arXiv preprint arXiv:1602.01024* (2016).
[27] Zhou Wang and Alan C Bovik. 2009. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE signal processing magazine* 26, 1 (2009), 98–117.
[28] Lior Wolf, Tal Hassner, and Itay Maoz. 2011. Face recognition in unconstrained videos with matched background similarity. In *IEEE CVPR*. 529–534.
[29] Rongkai Xia, Yan Pan, Lei Du, and Jian Yin. 2014. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *AAAI*, Vol. 28.
[30] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).
[31] Deyan Xie, Xiangdong Zhang, Quanxue Gao, Jiale Han, Song Xiao, and Xinbo Gao. 2019. Multiview clustering by joint latent representation and similarity learning. *IEEE TC* 50, 11 (2019), 4848–4854.
[32] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*. 478–487.
[33] Wei Xu, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *ACM SIGIR*. 267–273.
[34] Jianwei Yang, Devi Parikh, and Dhruv Batra. 2016. Joint unsupervised learning of deep representations and image clusters. In *IEEE CVPR*. 5147–5156.
[35] Yan Yang and Hao Wang. 2018. Multi-view Clustering: A Survey. *Big Data Mining and Analytics* 1, 2 (2018), 83–107.
[36] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. 2018. Binary multi-view clustering. *IEEE TPAMI* 41, 7 (2018), 1774–1782.
[37] Xiaofeng Zhu, Shichao Zhang, Rongyao Hu, Wei He, Cong Lei, and Pengfei Zhu. 2018. One-step multi-view spectral clustering. *IEEE TKDE* (2018).