End-to-End Adversarial-Attention Network for Multi-Modal Clustering

Runwu Zhou[1,2] Yi-Dong Shen [1*]
[1] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
{zhourw, ydshen}@ios.ac.cn

Abstract

Multi-modal clustering aims to cluster data into different groups by exploring complementary information from multiple modalities or views. Little work learns the deep fused representations of multiple modalities and simultaneously discovers the cluster structure with a discriminative loss. In this paper, we present an End-to-end Adversarial-attention network for Multi-modal Clustering (EAMC), where adversarial learning and attention mechanism are leveraged to align the latent feature distributions and quantify the importance of modalities respectively. To benefit from the joint training, we introduce a divergence-based clustering objective that not only encourages the separation and compactness of clusters but also enjoy a clear cluster structure by embedding the simplex geometry of the output space into the loss. The proposed network consists of modality-specific feature learning, modality fusion and cluster assignment three modules. It can be trained from scratch with batch-mode based optimization and avoid an autoencoder pre-training stage. Comprehensive experiments conducted on five real-world datasets show the superiority and effectiveness of the proposed clustering method.

1. Introduction

With the development of data collection techniques, multi-modal or view data has become an important part of current data resources in real-world applications. For example, in visual data, an image could be represented by different descriptor, such as SIFT, HoG and LBP, and a video contains audio signal and visual signal; in web news, a message could be delivered by pictures and texts. Although each modality has its own information and statistical properties, distinct modalities usually admit the same cluster structure. The rationale for using multi-modal data to learn the structured partition is that they can provide comprehensive estimation for the common pattern with the aid of the complementary information from modalities [33]. Recently, multi-modal clustering has gained significant momentum in machine learning and computer vision communities [3, 47].

A straightforward way to group this kind of data is firstly concatenating them into single-modal data and then resorting to single-modal clustering methods. However, this way can not guarantee good performance and even obtains worse results. As a consequence, the mainstream research is to learn low-dimensional latent representations such that the mutual agreement of modalities can be reached in the latent space. Recently, a variety of multi-modal clustering methods have been proposed, including CCA-based methods [7, 39], matrix factorization based methods [5, 42, 37], subspace learning based methods [44, 45, 48] and graph model based methods [32, 34]. Although these proposed methods have achieved promising results, they are greatly limited due to the use of shallow and linear embedding functions, which are not able to capture the nonlinear nature of complex data. To tackle this issue, some multiple kernel learning based methods [11, 27] have been proposed. However, it is difficult to select the proper kernel functions.

With the rapid development of deep neural network (DNN) models that are able to capture complex features in single-modal scenarios, such as image clustering, DNN has increasingly been exploited in multi-modal clustering task. Existing DNN-based multi-modal clustering methods fall into two categories. The first category regards multi-modal feature learning and cluster assignment as separated processes. The representative methods of this branch are DCCA [2] and DMSC [1]. DCCA first maximizes the correlation between the projected deep features of two views by CCA and then conducts the subsequent K-means clustering. DMSC takes convolutional neural networks for multi-modal subspace learning and next does spectral clustering based on the learned affinity graph. This kind of two-step learning strategy may disconnect closely related processes of feature learning and cluster assignment. The direct affect is that the learned representations can not friendly adapt to the predefined clustering algorithm. To close this gap, the other category unifies these two processes into joint opti-

*Corresponding author

mization step. DAMC [25] exemplifies this line of work. It works by pretraining multi-view autoencoder and then jointly optimizing the consensus cluster centroids, autoencoder networks and adversarial networks. Although DAMC has gained satisfactory results, it still faces some issues. On one hand, it equally treats each modality regardless of the quality difference among modalities, which makes it difficult to obtain optimal latent representations for clustering. On the other hand, the clustering loss used in this method overly relies on good initialization of pretaining stage. Moreover, it is difficult for this loss to ensure clear cluster structure since marginal samples are weakened and thus may not walk towards the correct clusters. On the whole, this line of research is in its infancy and at least two key problems are under explored: (1) How to learn the deep fused representations across multiple modalities? (2) What kind of loss function is suitable to train deep neural network for multi-modal clustering analysis?

In this paper, we propose an End-to-end Adversarial-attention Multi-modal Clustering (EAMC) method, which unifies multi-modal feature learning, modality fusion as well as clustering analysis into a joint process. The proposed method is built on the concepts of adversarial learning [14], attention mechanism [8] as well as information-theoretic divergence measures [17]. To be specific, we propose to align latent feature distribution of different modalities by introducing the adversarial regularizer. Through the adversarial process, modality invariance in the latent space can be reached more efficiently. We argue that a better alignment of modality distributions contributes to the subsequent fusion especially when the fused features are obtained by weighted average of latent features. Besides, we propose to quantify the importance of different modalities by introducing attention layer, which adaptively assigns the weight for each modality. Furthermore, we introduce a divergence-based clustering loss to guide the network training. The clustering loss we defined explicitly encourages the separation between clusters and the compactness within clusters, which are desirable properties to increase identifiability of clustering model. In addition, a precise geometry property of the output space induced by the softmax function is embedded into the Cauchy-Schwartz divergence to avert the degenerated structure of the clustering partition. It is also worth mentioning that the proposed clustering model can be trained from scratch without an autoencoder-based pretraining, compared to existing deep multi-modal clustering methods.

Figure 1 shows the overview of the proposed network architecture. In general, the proposed method consists of three main parts, i.e., modality-specific feature learning, modality fusion and cluster assignment. The modality-specific feature learning is designed to estimate the data similarity in the low-dimentional latent space, which also acts as feature encoders (or generators) to reveal the non-linearity of data. The modality fusion is constituted of modality alignment and modality-awareness modules. Concretely, a min-max game is played between a set of discriminators and generators in modality-alignment module to steer feature distribution learning. Meanwhile, three fully connected layers and a sigmoid layer are deployed in modality-awareness module to learn the weights of modalities. In tail, the cluster assignment layer made up of two fully connected layer and a softmax layer is added to conduct the network training with the defined loss. To sum up, the main contributions for multi-modal clustering community are as follows:

- A deep end-to-end multi-modal clustering method which unifies modality-specific feature learning, fusion and cluster assignment into a joint optimization procedure is proposed. Besides, for the first time adversarial learning and attention mechanism are simultaneously introduced for modality fusion process.
- A new discriminative clustering loss is defined to guide the network training. This loss explicitly encourages the separation and compactness of clusters and meanwhile ensures clear cluster structure by embedding the simplex geometry.
- Experimental results on five datasets convey the effectiveness and superiority of the proposed method.

2. Related Works

There are significant works on multi-modal clustering problem. From the perspective of representation learning, existing multi-modal clustering methods can be categorized into two groups, i.e., traditional and deep methods.

Traditional multi-modal clustering methods can be roughly divided into five streams. The methods in the first stream use non-negative matrix factorization technique to seek a common latent factor among multi-modal data [46, 42]. For instance, Cai et al. [5] formulated multi-modal clustering as the constrained matrix factorization problem with a shared clustering indicator matrix across different modalities. The methods in the second stream take self-representation way to characterize the relationships between the samples [44, 45, 6, 26]. A recent work [28] proposed to simultaneously learn a shared consistent representations and a set of view-specific representations for multi-modal subspace clustering. The methods in third stream exploit dimensionality reduction technique to firstly learn low-dimentional subspace and then conduct existing clustering algorithms to get the results [4, 7]. For this branch, canonical correlation analysis (CCA) [7] is a represntative method for multi-modal clustering which projects the multi-modal high dimensional data into a low-dimensional sub-

space by maximizing the correlation. The methods in the fourth stream exploit graph model for multi-modal clustering [32, 34]. The basic idea of this line is to find a consensus graph across multiple modalities and then uses graph-cut algorithms, e.g., spectral clustering, on the consensus graph to get clustering results. The limitation of the above methods is that they use shallow and linear embedding functions which can not reavel the nonlinear nature of complex data. The methods in the last stream draw support from kernel trick to address this issue [13, 24, 40, 11, 27]. Usually some predefined kernel functions, e.g. Gaussian kernel, are required to deal with different modalities. These kernel functions are then combined either linearly or nonlinearly to arrive at a consensus kernel. The difficulty of this stream lies in the choice of kernel functions.

Deep neural networks have increasingly been exploited in multi-modal clustering issue due to powerful feature transformation ability. In the early stage, Ngiam et al. [30] took deep auto-encoder network architecture to learn the common representations of multi-modal data and achieved superior performance in speech and vision tasks. Later, Andrew et al. [2] proposed a deep extension of CCA (DCCA) to learn the common representations by maximizing the correlation with CCA based on extracted deep features. Recently, Wang et al. [39] developed a new CCA variant by merging DCCA and autoencoder. Later, Abavisani et al. [1] introduced deep multi-modal subspace clustering network to find a shared affinity across all modalities. A disadvantage of the above deep models is that two closely related tasks of feature learning and clustering are disconnected. To make these two tasks benefit from each other, Li et al. [25] proposed a joint learning framework (DAMC) for multi-modal clustering and achieved current state-of-the-art performance.

The method proposed in this paper falls into the category of joint learning methods. Our method is inspired, on the one hand, by the ideas in traditional multi-modal clustering methods, especially regarding weight learning aspect [31, 43, 38], which have been found extremely influential for clustering results. On the other hand, our method is also inspired by the success of adversarial learning in many tasks, such as cross-modal retrieval [36], domain adaptation [9]. Furthermore, our method is especially inspired by the superiority of divergence-based clustering for single modal data [17, 20, 35].

3. The Proposed Method

Consider the problem of clustering a set of n data points consisting of V modalities $\mathcal{D} = \{\mathbf{X}^1, ...\mathbf{X}^v, ..., \mathbf{X}^V\}$ into c clusters, where $\mathbf{X}^v \in \mathcal{R}^{d_v \times n}$ denotes the samples of the dimension d_v from the v-th modality. We build an end-to-end adversarial-attention clustering network to make it. In the following, we first introduce the proposed network architecture and then describe the defined loss function.

3.1. Network Architecture

The proposed network architecture consists of modality-specific feature learning, modality fusion and cluster assignment, which is illustrated in Figure 1.

(A) Modality-Specific Feature Learning

Different statistical properties of multi-modal data hint it is rather difficult to fuse different modalities in the data space. In light of this, we design a modality-specific feature learning module to transform the data into low-dimentional latent space. This module performs the main task of feature learning. It also has a objective to confuse the discriminator, which we will discuss later. To be concrete, for the v-th modality, we firstly encode the corresponding latent features as $\mathbf{H}^v = E_v(\mathbf{X}^v; \theta_e^v)$, where $E_v(.)$ refers to the v-th modality's encoder parameterized by θ_e^v. Then based on $\mathbf{H}^v$, we can estimate data metric, e.g., Gaussian metric, of data in latent space. Formally, it can be written as $\mathbf{K}_{ij}^v = exp(-||\mathbf{h}_i^v - \mathbf{h}_j^v||_2/2\sigma^2)$. Here, $\mathbf{h}_i^v$ denotes the i-th column of $\mathbf{H}^v$ (i.e., i-th sample) and σ represents the bandwidth. Note that we constrain encoder networks with random i.i.d Gaussian weights to avoid degenerated metric structure of data, which is different from existing deep multi-modal clustering methods that leverage the decoder network for the purpose. This design choice takes the inspiration from the recent advanced work in the theory of neural networks [12]. It showed that the metric structure of data can be preserved by DNN with random i.i.d Gaussian weights when the intrinsic dimensionality of the data is proportional to the network width. In the experiment, this can be met since [12] proved that the intrinsic dimensionality of the data does not increase as the data propagate through the network.

(B) Modality Fusion

Modality fusion module is designed to fuse diverse information of different modalities for comprehensive estimation. In our model, this module is made of modality alignment and modality-awareness submodules.

Modality alignment submodule serves for aligning the latent feature distributions of modalities. It consists of $V-1$ discriminators, each of which is three fully connected layers. Specifically, taking the first modality as an anchor, we assign a discriminator between the first modality and one of the rest in pairwise fashion. For each latent feature $\mathbf{H}_v (v = 2, 3, ..., V)$ drawn from the distribution p_v, the discriminator D_v parameterized by θ_d^v aims to verify whether its real data $\mathbf{h}_i^1 \in \mathbf{H}^1$ and fake data $\widetilde{\mathbf{h}_i^v} \in \mathbf{H}^v$ belong to the same distribution. In this process, the discriminator network D_v is optimized in alternating manner with the encoder network E_v to solve adversarial min-max problem [14]. By this way, the discriminator networks can guide encoder networks to learn the same latent feature distribution. Note

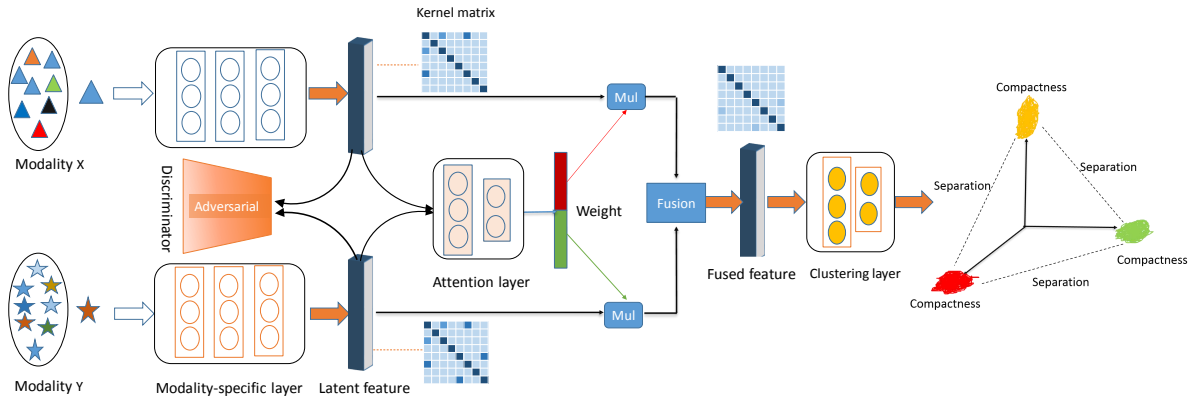

Figure 1. Illustration of the proposed EAMC network (here we take two modalities X and Y as example). EAMC consists of modality-specific feature learning module, modality fusion mudule (modality alignment and modality-awareness) and cluster assignment module. Modality-specific feature learning is to learn non-linear property of data and estimate data similarity in latent space. Modality fusion module aims to align the feature distributions and quantify the importance of modalities. Finally, a cluster assignment layer is applied to guide the network training by a discriminative loss.

that considering all possible combinations (up to 2^V) will dramatically increase the burden of network training.

Modality-awareness submodule is introduced to learn the weights for different modalities, the input of which is the concatenated features $\mathbf{h}$ and the output of which is a V-dimensional vector $\mathbf{w}$. In general, it is composed of three fully connected layers and a softmax layer. We describe the process with the following formulas:

$$\mathbf{h} = [\mathbf{h}^1, \mathbf{h}^2, ..., \mathbf{h}^V], \tag{1}$$

$$\mathbf{act} = \text{FCs}(\mathbf{h}), \tag{2}$$

$$\mathbf{e} = \text{Softmax}(\text{sigmoid}(\mathbf{act})/\tau), \tag{3}$$

$$\mathbf{w} = \text{Mean}(\mathbf{e}, \text{dim=0}) \tag{4}$$

where $[\cdot]$ denotes the concatenation operator; $\text{FCs}(\cdot)$ represents 3 fully connected layers; τ is a calibration factor. Sigmoid function together with calibration factor can be regarded as a trick to avoid assigning close-to-one score to the most related modality. For the sake of simplicity, the parameters in this module are denoted as θ_a.

At this time, we can get the fused representations of the sample with the formulas: $\mathbf{h}_f = \sum_v \mathbf{w}_v \mathbf{h}^v$. Then, $\mathbf{h}_f$ is fed into the clustering layer to get soft clustering assignment.

(C) Cluster Assignment

In order to benefit from joint learning methods, we deploy a clustering layer parameterized by θ_c in the network. The clustering layer is stacked on the top of fusion layer, which consists of two fully connected layer and a softmax layer. The softmax layer output the soft cluster membership matrix $\mathbf{A} = [\alpha_{qi}]$, with elements $\alpha_{qi} \in (0, 1)$ that represents the crisp cluster assignment of data point q to cluster C_i . We then use the defined loss to guide network training.

3.2. Loss Function

1) Fusion Loss. In our model, the min-max game is played between the generators (encoders) and discriminators to steer feature distribution learning towards the first modality. The corresponding optimization objective for this purpose can be expressed as:

$$\mathcal{L}_{adv} = \min_{\theta_e^v} \max_{\theta_d^v} \sum_{v=2}^{V} \mathbb{E}_{h^1 \sim p_1}[logD_v(h^1)] + \mathbb{E}_{h^v \sim p_v}[log(1 - D_v(h^v))] \tag{5}$$

What's more, in order to make the metric structures of different modalities reach the mutual agreements, inspired by the success of companion loss [23] in supervised deep models, we impose the following loss on fusion module:

$$\mathcal{L}_{att} = ||\mathbf{K}^f - \mathbf{K}^c||_F^2 \tag{6}$$

where $\mathbf{K}^f$ is computed based on the fused features with Gaussian kernel and $\mathbf{K}^c = \sum_v \mathbf{w}_v \mathbf{K}^v$. The extra affect of (6) is that the weight is further considered in metric level such that the fused results are more reliable.

2) Clustering Loss. In order to learn a good partition structure, recent advances [25, 41] usually use Kullback-Leibler (KL) divergence based loss to guide clustering process. It works by emphasizing on data points assigned with high confidence. The way like this does not necessarily enforces cluster compactness due to the neglect for the marginal samples. In this section, we introduce a new clustering loss based on Cauchy-Schwarz divergence to alleviate this issue. The introduced clustering loss encourages the separation between clusters and the compactness within clusters. Meanwhile, it also explicitly exploits the geometry structure of the output space during the optimization.

Here we firstly recap the definition of multiple-pdf generalization of the Cauchy-Schwartz (CS) divergence[17]:

$$\mathcal{D}_{sc} = -\log\Big(\frac{1}{k}\sum_{i=1}^{k-1}\sum_{j>i}\frac{\mathbb{E}_{\mathbf{h}\sim p_i}(p_j(\mathbf{h}))}{\sqrt{\mathbb{E}_{\mathbf{h}\sim p_i}(p_i(\mathbf{h}))\mathbb{E}_{\mathbf{h}\sim p_j}(p_j(\mathbf{h}))}}\Big) \tag{7}$$

where k is the number of distributions, p_i and p_j respectively denotes probability density functions (pdf) of the cluster C_i and C_j. A large divergence would lead to well separated and compact clusters. According to a data-driven approach [20], maximizing (7) is in practice equivalent to minimizing the following formula:

$$\mathcal{D}_{sc} = \frac{1}{k}\sum_{i=1}^{k-1}\sum_{j>i}\frac{\boldsymbol{\alpha}_i^T\mathbf{K}\boldsymbol{\alpha}_j}{\sqrt{\boldsymbol{\alpha}_i^T\mathbf{K}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j^T\mathbf{K}\boldsymbol{\alpha}_j}} \tag{8}$$

where $\mathbf{K}$ is data metric matrix based on Gaussian kernel. The vectors $\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, ..., \boldsymbol{\alpha_k}$ denote the columns of the hard cluster assignment matrix $\mathbf{A} \in \mathbb{R}^{n\times k}$. In our architecture, we relax the hard membership to soft one in order to preserve differentiability of the loss.

Furthermore, to avoid a degenerated clustering partition, we exploit the output space property, i.e., a simplex in $\mathbb{R}^k$, induced by the softmax activation to enforce the closeness of the output to a corner of the simplex. Concretely, we intergrate this geometry structure into CS divergence by the following form:

$$\mathcal{D}_{sim} = \frac{1}{k}\sum_{i=1}^{k-1}\sum_{j>i}\frac{\boldsymbol{\beta}_i^T\mathbf{K}\boldsymbol{\beta}_j}{\sqrt{\boldsymbol{\beta}_i^T\mathbf{K}\boldsymbol{\beta}_i\boldsymbol{\beta}_j^T\mathbf{K}\boldsymbol{\beta}_j}} \tag{9}$$

where $\boldsymbol{\beta}_i$, $\boldsymbol{\beta}_j$ are the i-th, j-th column of the matrix $\mathbf{B} = [\beta_{qi}]$ with $\beta_{qi} = exp(-||\boldsymbol{\alpha}_q - \mathbf{e}_i||)$. Here $\mathbf{e}_i$ denotes the i-th corner of the simplex. By this way, the cluster assignment vectors would be compactly centered around distinct simplex corners. In the experiments, $\mathbf{K}$ is replaced with $\mathbf{K}^f$.

Lastly, we hope that the clusters are orthogonal in n-dimentional observation space. Mathematically, it can be formulated as

$$\mathcal{D}_{reg} = triu(\mathbf{A}^T\mathbf{A}) \tag{10}$$

where triu$(\cdot)$ denotes the sum of the strictly upper triangular elements of its argument. Now, we can write the total clustering loss as

$$\mathcal{L}_c = \mathcal{D}_{sc} + \mathcal{D}_{sim} + \mathcal{D}_{reg} \tag{11}$$

3.3. Optimization Training

We present the detailed optimization steps in Algorithm 1. From the perspective of adversarial optimization, the proposed EAMC is conducted by alternately optimizing the following processes:

$$(\hat{\theta}_e^v, \hat{\theta}_a, \hat{\theta}_c) = \arg\min_{\theta_e^v,\theta_a,\theta_c}(\mathcal{L}_c + \mathcal{L}_{att} - \gamma\mathcal{L}_{adv}) \tag{12}$$

$$\hat{\theta}_d^v = \arg\max_{\theta_d^v}(\mathcal{L}_c + \mathcal{L}_{att} - \gamma\mathcal{L}_{adv}) \tag{13}$$

Algorithm 1 Pseudocode of optimizing our EAMC

Initialization: Batch multi-modal data (of size m) $\mathcal{D}_b = \{\mathbf{X}_b^1, \mathbf{X}_b^2, ..., \mathbf{X}_b^V\} \in \mathcal{D}$;
The hyperparameters γ and t;
Initialize encoder networks with random i.i.d Gaussian weights in order to preserve the metric structure [12];

Update until convergence:

for t steps **do**
  update parameters θ_e^v, θ_a and $\theta_c (v = 1, 2, ..., V)$ by descending their stochastic gradients:
    $\theta_e^v \leftarrow \theta_e^v$ - $\eta \cdot \nabla_{\theta_e^v}\frac{1}{m}(\mathcal{L}_c + \mathcal{L}_{att} - \gamma\mathcal{L}_{adv})$
    $\theta_a \leftarrow \theta_a$ - $\eta \cdot \nabla_{\theta_a}\frac{1}{m}(\mathcal{L}_c + \mathcal{L}_{att} - \gamma\mathcal{L}_{adv})$
    $\theta_c \leftarrow \theta_c$ - $\eta \cdot \nabla_{\theta_c}\frac{1}{m}(\mathcal{L}_c + \mathcal{L}_{att} - \gamma\mathcal{L}_{adv})$
end for
update parameters θ_d^v of the discriminator by ascending its stochastic gradient:
    $\theta_d^v \leftarrow \theta_d^v$ + $\eta \cdot \nabla_{\theta_d^v}\frac{1}{m}(\mathcal{L}_c + \mathcal{L}_{att} - \gamma\mathcal{L}_{adv})$
return Cluster assignment matrix $\mathbf{A}$;

4. Experiments

4.1. Experimental Setup

Datasets *NUS-WIDE-C5(NWC)*: A image-text dataset consists of 4,000 objects for 5 classes (bird, food, sun, tower, toy). Each class has 800 objects which is represented by a 500-dimensional visual codeword vector and 1000-dimensional annotation vector. *SentencesNYUv2 (RGB-D)*: A dataset includes 1,449 images with 13 indoor scenes. Every image is captioned with a paragraph which describes the content of the image. We use ResNet-50, pretrained on ImageNet, to extract 2048 dimentional image features and doc2vec, pretrained on Wikipedia via skip-gram, to extract 300 dimentional text features. *Pascal VOC*: A dataset includes 9,963 image-text pairs with 20 classes. Each image is represented by a 512-D Gist Feature vector and each text is represented as 399-dimensional word frequency count. We pick 5,649 images with only one object in our experiment. *The Columbia Consumer Video (CCV)*: A dataset contains 9,317 YouTube videos with 20 diverse semantic categories. We use the subset (6773 videos) of CCV provided by [18], along with three hand-crafted features: STIP features with 5,000 dimensional Bag-of-Words (BoWs) representation, SIFT features extracted every two

seconds with 5,000 dimensional BoWs representation, and MFCC features with 4,000 dimensional BoWs representation. *MNIST*: A large-scale handwritten digit dataset includes 70,000 samples with 28 $\times$ 28 pixels. The first view is the original gray images, and the other is given by images only highlighting the digit edge. Table 1 provides a brief description of each dataset.

Dataset	type	#sample	#modal	#class
NWC	image-text	4,000	2	5
RGB-D	image-text	1,449	2	13
VOC	image-text	5,649	2	20
CCV	video	6,773	3	20
MNIST	digit	70,000	2	10

Table 1. Dataset Description

Evaluation Metrics The clustering performance is measured using two standard evaluation matrices, i.e., Accuracy (ACC) and Normalized Mutual Information (NMI). For the two metrics, higher value indicates better performance. More details about the two metrics refer to [22].

Implementation Details The proposed network architecture is trained with the PyTorch platform. We use a common architecture for EAMC in order to provide a practical method for real-world datasets. For all types of data, we firstly transform them into vectorial representation and then feed them into the network. In the experiments, we use the Adam solver [21] with a batch size of 100. Training is performed with learning rate 10^{-3} for the encoder and discriminator networks, 10^{-4} for the attention layer and 10^{-5} for the clustering layer. For each iteration, we reshuffle the ordering of the mini-batches. Weights of the network are initialized following [15]. The kernel width, σ, is set to 15% of the median pairwise distance between the latent representations within each batch following [16]. To increase the models' robustness, batch-normalization is applied before the softmax output. As unsupervised deep models easily get stuck in local minima, we run EAMC for 20 runs and report the accuracy of the run with the lowest clustering loss.

Baseline Models To evaluate the performance of our method, we compare it with the following methods:

(A) Spectral clustering (SC). The standard spectral clustering algorithm [29] is conducted on every modality and the concatenated modality.

(B) Traditional Methods. 1) RMKMC: Robust multi-view k-means clustering (RMKMC) [5] searches a consensus cluster indicator across multiple views; 2) tRLMvc: Tensor-based representation learning multi-view clustering (tRLMvc) [10] unifies the self-expressive tensor learning and low-dimensional representation learning together to capture the essential structure hidden in the multi-view data; 3) CSMCS: Consistent and specific multi-view subspace clustering (CSMCS) [28] formulates the multi-view self-representation property using a shared consistent representation and a set of specific representations; 4) WMSC: Weighted multi-view spectral clustering (WMSC) [49] employs spectral perturbation theory to model the weights of modalities; 5) MCGC: Multi-view consusens graph clustering (MCGC) [19] learns a consensus graph with minimizing disagreement between different views and constraining the rank of the Laplacian matrix.

(C) Deep Methods. 1) DCCA: Deep canonical correlation analysis (DCCA) [2] learns nonlinear transformations of two views such that the extracted features are highly linearly correlated; 2) DMSC: Deep multimodal subspace clustering (DMSC) [1] presents convolutional neural network based approaches for unsupervised multimodal subspace clustering; 3) DAMC: Deep adversarial multi-view clustering (DAMC) [25] adopts deep auto-encoders to learn latent representations shared by multiple views and meanwhile leverages adversarial training to further capture the data distribution.

The default parameters of each compared method are adopted in our experiments. For all these compared methods, we run each method 10 times and report the average performance. For the postprocessing methods (CSMCS, tRLMvc, WMSC, DCCA and DMSC), we run K-means clustering 20 runs and report the result with the minimum loss. Since CCA-based methods (CCA and DCCA) can only deal with two modalities, we choose the best two modalities on CCV dataset according to their performance.

4.2. Performance Evaluation

Compared with Baselines Experimental results are shown in Table 2. It can be seen that the clustering results from multi-modal clustering methods including traditional and deep methods significantly outperform that from the single-modal (based on only one or concatenated modality), which demonstrates the necessity of fusing multi-modal information for clustering. Compared with five traditional clustering methods, EAMC surpasses them with a large margin. For instance, on NWC, our model has a growth by (94.5-87.3) 7.2%, (93.7-86.2) 7.5% and (95.2-87.6) 7.6% against the second best method in terms of ACC, NMI and Purity. The pivotal reason behind this is that traditional methods are greatly limited by using shallow and linear embedding functions, which are not able to capture the complex property of real-world data. Besides, compared with deep models, our model also shows clear advantage. In particular, our method outperforms joint learning method DAMC with a clear improvement on all four datasets. We attribute this success to feature distribution alignment and weight learning among modalities.

Clustering on Large-scale Dataset In order to show our model is applicable on the large-scale dataset, we have conducted the experiments on MNIST dataset. The compared

Dataset	NWC			RGB-D			VOC			CCV		
Metric	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity	ACC	NMI	Purity
SC(1)	0.712	0.768	0.747	0.334	0.297	0.347	0.384	0.392	0.379	0.102	0.005	0.104
SC(2)	0.647	0.689	0.699	0.297	0.305	0.326	0.402	0.411	0.395	0.188	0.173	0.213
SC(3)	-	-	-	-	-	-	-	-	-	0.113	0.008	0.109
SC(con)	0.652	0.673	0.686	0.312	0.286	0.320	0.372	0.387	0.382	0.093	0.074	0.102
RMKMC	0.784	0.793	0.791	0.379	0.398	0.397	0.458	0.469	0.473	0.176	0.165	0.186
tRLMvc	0.873	0.849	0.869	0.445	0.439	0.460	0.534	0.547	0.556	0.212	0.226	0.231
CSMCS	0.824	0.813	0.829	0.392	0.414	0.426	0.488	0.496	0.517	0.194	0.186	0.198
WMSC	0.798	0.787	0.816	0.408	0.425	0.420	0.471	0.462	0.477	0.205	0.196	0.208
MCGC	0.853	0.862	0.876	0.438	0.447	0.453	0.527	0.546	0.539	0.224	0.216	0.240
DCCA	0.784	0.798	0.809	0.355	0.362	0.374	0.397	0.425	0.433	0.173	0.182	0.186
DMSC	0.877	0.864	0.876	0.419	0.426	0.433	0.541	0.538	0.566	0.183	0.194	0.196
DAMC	0.891	0.914	0.916	0.463	0.475	0.481	0.560	0.552	0.583	0.243	0.231	0.264
EAMC	**0.945**	**0.937**	**0.952**	**0.497**	**0.499**	**0.511**	**0.607**	**0.615**	**0.628**	**0.261**	**0.266**	**0.271**

Table 2. Clutering results on NWC, RGB-D, VOC and CCV datasets.

methods include three deep baseline models, i.e., DCCA, DMSC and DAMC. The other methods are not scalable on this dataset due to their optimization methods and the limited memory. Benefiting from the architecture design and loss function, our model is able to support batch-mode based optimization and thus easily addresses the large-scale multi-modal clustering issue. As shown in Table 3, EAMC clearly outperforms other deep models in ACC and NMI, which validates the effectiveness of the proposed model on the large-scale dataset.

Model	ACC	NMI	Purity
DCCA	0.476	0.443	0.492
DMSC	0.653	0.614	0.644
DAMC	0.646	0.594	**0.657**
EAMC	**0.668**	**0.628**	0.651

Table 3. Clustering result on large-scale MNIST dataset.

4.3. Further Evaluation

Component Study We train three variants to examine the effect of adversarial and attention components: (1) EAMC$_{att}$ denotes the network which is obtained by removing adversarial module in EAMC: (2) EAMC$_{adv}$ denotes the network which is obtained by removing attention module in EAMC; (3) EAMC$_{none}$ denotes the network which is obtained by removing both adversarial and attention modules in EAMC. After removing the attention layers, we assign the equal weight (i.e., $\mathbf{w}_v = \frac{1}{V}$) for each modality. Table 4 shows the experimental results on NWC dataset. Here some important observations can made as follows. Firstly, it can be seen that EAMC$_{att}$ and EAMC$_{adv}$ outperform EAMC$_{none}$ with a clear improvement. Additionally, EAMC further improve the performace compared with three variants. These results convey that adversarial and attention components are key technical choice for multi-modal clustering.

Model	ACC	NMI	Purity
EAMC$_{att}$	0.921	0.917	0.932
EAMC$_{adv}$	0.908	0.896	0.903
EAMC$_{none}$	0.871	0.884	0.892
EAMC	**0.945**	**0.937**	**0.952**

Table 4. Component study on NWC dataset.

Loss Analysis We empirically analyze the clustering loss to evaluate the influence of different terms. The accuracy results for the NWC and RGB-D datasets are reported in Table 5. First of all, it is clearly observed that combining the term $\mathcal{D}_{sc}$ with $\mathcal{D}_{sim}$ greatly boost the performance. In addition, by using three terms together, the performance can be further improved.

Loss	NWC	RGB-D
$\mathcal{D}_{sc}$	0.836	0.364
$\mathcal{D}_{sim}$	0.852	0.379
$\mathcal{D}_{sc} + \mathcal{D}_{sim}$	0.918	0.437
$\mathcal{D}_{sc} + \mathcal{D}_{reg}$	0.877	0.426
$\mathcal{D}_{sim} + \mathcal{D}_{reg}$	0.898	0.412
$\mathcal{D}_{sc} + \mathcal{D}_{sim} + \mathcal{D}_{reg}$	**0.945**	**0.497**

Table 5. Clustering loss analysis on NWC and RGB-D datasets

Weight Score Different modalities usually make distinct contributions to the final clustering results. To clearly see this fact, we report the weight score in Table 6. For example, on NWC, the weight of annotation vector is larger than that of codeword vector, which reflects the modality of annotation vector would provide more useful information for clustering. On MNIST dataset, EAMC considers the edge modality palys more important role for clustering. The similar phenomena can be observed in the remaining three datasets. The results in Table 6 are in accordance with

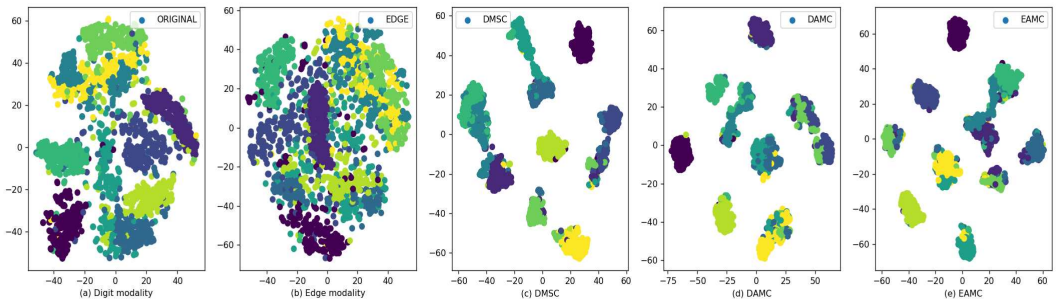

Figure 2. Visualization of original pixel features for each modality and the fused features obtained through competitive baselines with t-SNE on the MNIST dataset. (a) Original digit image features of the first modality, (b) Edge image of the second modality, (c) DMSC, (d) DAMC, and (e) EAMC.

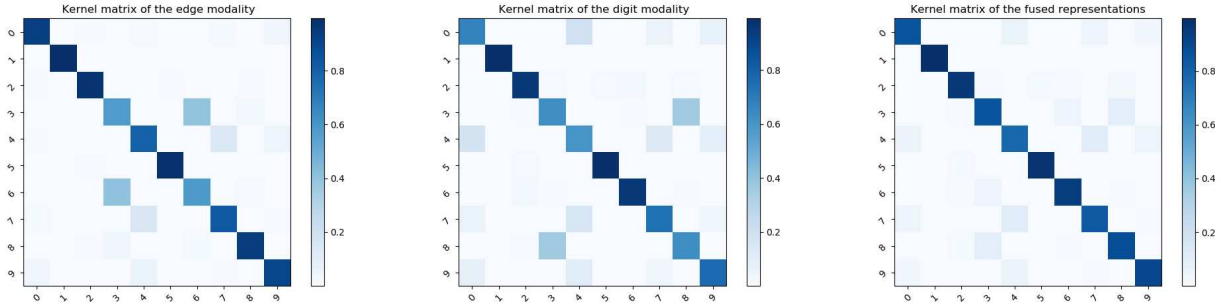

Figure 3. Visualization of the kernel matrix computed over the latent features on the MNIST dataset. From the left to right, (a) kernel matrix of the edge modality, (b) kernel matrix of the digit modality, (c) kernel matrix of the fused representations.

the idea that different modalities usually have a distinct contributions to clustering results.

Dataset	modal-1	modal-2	modal-3	Relation
NWC	0.438	0.562	-	2 >1
RGB-D	0.467	0.533	-	2 >1
VOC	0.483	0.517	-	2 >1
CCV	0.257	0.384	0.359	2>3>1
MNIST	0.477	0.523	-	2>1

Table 6. Weight score of different modalities for clustering on five datasets. The symbol '>' denotes the degree of importance.

Visualization To further evaluate the advantage of the proposed model over other deep models, we provide a t-SNE visualization for latent features of clustering layer on MNIST dataset. Two deep models, i.e., DMSC and DAMC, are selected for comparison. We randomly pick 2,000 samples and visualize two-dimensional embedded features of the fused representations. The visualization results are shown in Figure 2. It is clear that EAMC gives a more clear and compact cluster structure than the baseline models. Furthermore, we also provide a visualization of kernel matrix computed over the latent representations. It can be seen from Figure 3 that the kernel matrix of the fused representations reflects a more accurate block structure compared with the kernel matrices of the edge and digit modalities computed over the respective latent space.

5. Conclusion

In this paper, we propose an end-to-end adversarial-attention network for multi-modal clustering (EAMC). The proposed method exploits adversarial learning and attention mechanism to align the latent feature distributions and quantify the importance of modalities respectively. Besides, a discriminative clustering loss that not only encourages the separatin and compactness of the clusters but also enjoy a clear cluster structure is introduced to support end-to-end training. The proposed network consisting of modality-specific feature learning, modality fusion and cluster assignment three modules can be trained from scratch without an extra initialization component. Experimental results on five real-world datasets show the superiority and effectiveness of the proposed method.

Acknowledgements This work is supported in part by China National 973 program 2014CB340301 and NSFC grant 6197023605.

References

[1] Mahdi Abavisani and Vishal M Patel. Deep multimodal subspace clustering networks. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1601–1614, 2018.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255, 2013.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[4] Matthew B Blaschko and Christoph H Lampert. Correlational spectral clustering. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[5] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *Twenty-Third International Joint conference on artificial intelligence*, 2013.

[6] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–594, 2015.

[7] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM, 2009.

[8] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[9] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570, 2018.

[10] Miaomiao Cheng, Liping Jing, and Michael K Ng. Tensor-based low-dimensional representation learning for multi-view clustering. *IEEE Transactions on Image Processing*, 28(5):2399–2414, 2018.

[11] Liang Du, Peng Zhou, Lei Shi, Hanmo Wang, Mingyu Fan, Wenjian Wang, and Yi-Dong Shen. Robust multiple kernel k-means using l21-norm. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[12] Raja Giryes, Guillermo Sapiro, and Alex M Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Transactions on Signal Processing*, 64(13):3444–3457, 2016.

[13] Mehmet Gönen and Adam A Margolin. Localized data fusion for kernel k-means clustering with application to cancer biology. In *Advances in Neural Information Processing Systems*, pages 1305–1313, 2014.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[16] Robert Jenssen. Kernel entropy component analysis. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):847–860, 2009.

[17] Robert Jenssen, Jose C Principe, Deniz Erdogmus, and Torbjørn Eltoft. The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006.

[18] Yu Gang Jiang, Guangnan Ye, Shih Fu Chang, Daniel P. W. Ellis, and Alexander C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Acm International Conference on Multimedia Retrieval*, 2011.

[19] Zhan K, Nie F, Wang J, and Yang Y. Multiview consensus graph clustering. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 2019.

[20] Michael Kampffmeyer, FM Bianchi, L Livi, A-B Salberg, R Jenssen, et al. Deep divergence-based clustering. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2017.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *Advances in neural information processing systems*, pages 1413–1421, 2011.

[23] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570, 2015.

[24] Miaomiao Li, Xinwang Liu, Lei Wang, Yong Dou, Jianping Yin, and En Zhu. Multiple kernel clustering with local kernel alignment maximization. 2016.

[25] Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, and Zhaohua Yang. Deep adversarial multi-view clustering netwrok. In *IJCAI*, pages 2952–2958, 2019.

[26] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012.

[27] Xinwang Liu, Yong Dou, Jianping Yin, Lei Wang, and En Zhu. Multiple kernel k-means clustering with matrix-induced regularization. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[28] Shirui Luo, Changqing Zhang, Wei Zhang, and Xiaochun Cao. Consistent and specific multi-view subspace clustering. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[29] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings*

of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, 2001.

[30] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[31] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[32] Feiping Nie, Jing Li, Xuelong Li, et al. Self-weighted multiview clustering with multiple graphs. In *IJCAI*, pages 2564–2570, 2017.

[33] Sarah Rastegar, Mahdieh Soleymani, Hamid R Rabiee, and Seyed Mohsen Shojaee. Mdl-cw: A multimodal deep learning framework with cross weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2601–2609, 2016.

[34] Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. Marginalized multiview ensemble clustering. *IEEE transactions on neural networks and learning systems*, 2019.

[35] Vidar V Vikjord and Robert Jenssen. Information theoretic clustering using a k-nearest neighbors approach. *Pattern Recognition*, 47(9):3070–3081, 2014.

[36] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 154–162. ACM, 2017.

[37] Jing Wang, Feng Tian, Hongchuan Yu, Chang Hong Liu, Kun Zhan, and Xiao Wang. Diverse non-negative matrix factorization for multiview data representation. *IEEE transactions on cybernetics*, 48(9):2620–2632, 2017.

[38] Rong Wang, Feiping Nie, Zhen Wang, Haojie Hu, and Xuelong Li. Parameter-free weighted multi-view projected clustering with structured graph learning. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[39] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015.

[40] Yueqing Wang, Xinwang Liu, Yong Dou, Qi Lv, and Yao Lu. Multiple kernel learning with hybrid kernel alignment maximization. *Pattern Recognition*, 70:104–111, 2017.

[41] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487, 2016.

[42] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view self-paced learning for clustering. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[43] Jinglin Xu, Junwei Han, and Feiping Nie. Discriminatively embedded k-means for multi-view clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2016.

[44] Zhiyong Yang, Qianqian Xu, Weigang Zhang, Xiaochun Cao, and Qingming Huang. Split multiplicative multi-view subspace clustering. *IEEE Transactions on Image Processing*, 2019.

[45] Ming Yin, Junbin Gao, Shengli Xie, and Yi Guo. Multi-view subspace clustering via tensorial t-product representation. *IEEE Transactions on Neural Networks and Learning Systems*, 30(3):851–864, 2018.

[46] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[47] Peng Zhou, Liang Du, Lei Shi, Hanmo Wang, and Yi-Dong Shen. Recovery of corrupted multiple kernels for clustering. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[48] Tao Zhou, Changqing Zhang, Xi Peng, Harish Bhaskar, and Jie Yang. Dual shared-specific multiview subspace clustering. *IEEE transactions on cybernetics*, 2019.

[49] Linlin Zong, Xianchao Zhang, Xinyue Liu, and Hong Yu. Weighted multi-view spectral clustering based on spectral perturbation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.