Deep Multi-view Learning Methods: A Review

Xiaoqiang Yan[a], Shizhe Hu[a], Yiqiao Mao[a], Yangdong Ye[a,*], Hui Yu[b,*]

[a]*School of Information Engineering, Zhengzhou University, Zhengzhou, 450052, China*
[b]*School of Creative Technologies, University of Portsmouth, PO1 2DJ, United Kingdom*

Abstract

Multi-view learning (MVL) has attracted increasing attention and achieved great practical success by exploiting complementary information of multiple features or modalities. Recently, due to the remarkable performance of deep models, deep MVL has been adopted in many domains, such as machine learning, artificial intelligence and computer vision. This paper presents a comprehensive review on deep MVL from the following two perspectives: MVL methods in deep learning scope and deep MVL extensions of traditional methods. Specifically, we first review the representative MVL methods in the scope of deep learning, such as multi-view auto-encoder, conventional neural networks and deep brief networks. Then, we investigate the advancements of the MVL mechanism when traditional learning methods meet deep learning models, such as deep multi-view canonical correlation analysis, matrix factorization and information bottleneck. Moreover, we also summarize the main applications, widely-used datasets and performance comparison in the domain of deep MVL. Finally, we attempt to identify some open challenges to inform future research directions.

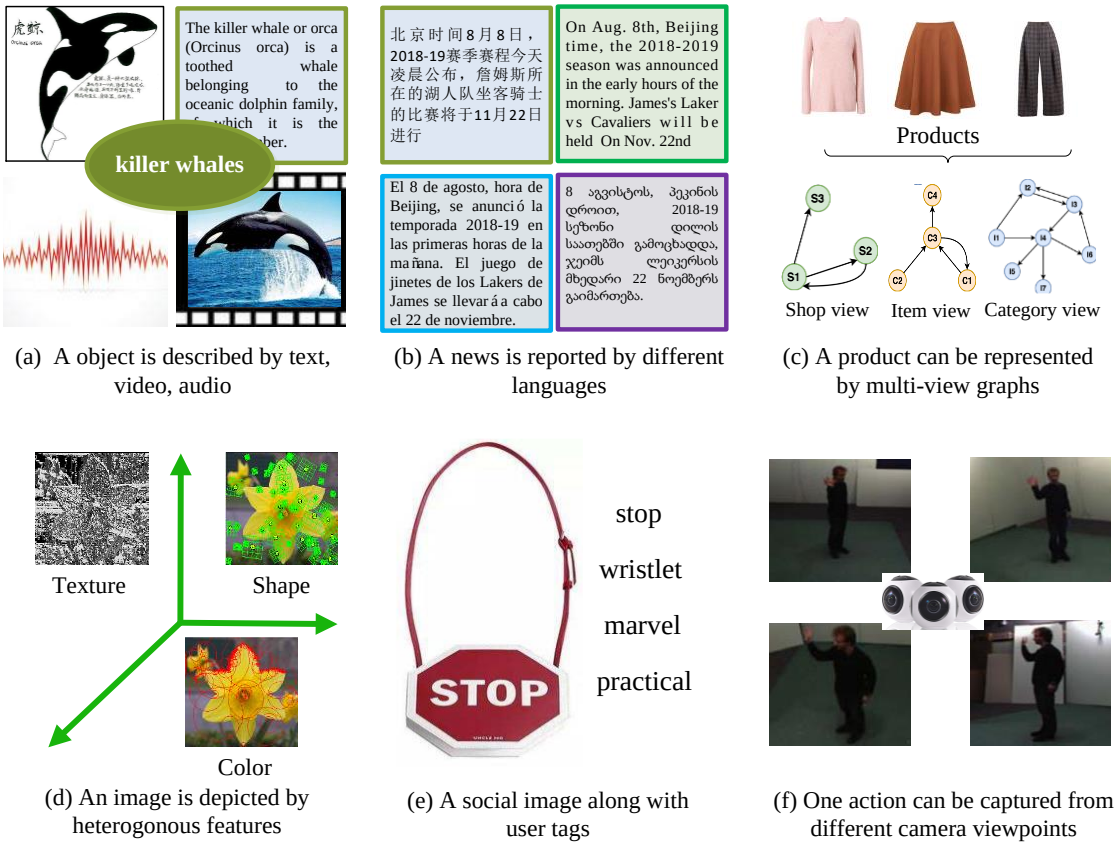
Keywords: Deep multi-view learning, deep neural networks, representation learning, statistical learning survey

1. Introduction

In recent decades, multi-view data has become one of the main data types on the Internet as its volume increases explosively in the domain of video surveillance [1, 2, 3], entertainment media [4, 5, 6], social network [7] and medical detection [8, 9], etc. Basically, multi-view data refer to the data captured from different modalities, sources, spaces and other forms, but with similar high-level semantics. As shown in Figure 1, an object can be described in forms of text, video, audio; An event is usually reported in different languages; A product can be represented by multiple graphs; A realistic image can be described by different visual features; A social image contains visual information and user tags; A specific human action can be captured by different cameras from various viewpoints. Although these views often represent diverse and complementary information of the same data, directly integrating them together does not obtain consistently satisfactory performance due to the biases between multiple views. Therefore, how to integrate multiple views properly is a central problem, which is also the objective of multi-view learning.

Multi-view learning (MVL) aims to learn the common feature spaces or shared patterns by combining multiple distinct features or data sources [10]. In the last decades, MVL has gained significant momentum in machine learning and computer vision communities [11, 12, 13, 14, 15] and inspired many promising algorithms, such as co-training mechanism [16], subspace learning methods [17] and multiple kernel learning (MKL) [18]. One of the most prevalent MVL approaches is to map multi-view data into a common feature space maximizing the mutual agreement of multiple views [19, 20, 21, 22, 7]. In this research direction, the early and representative one is canonical correlation analysis (CCA) [11], which is a statistical method that searches the linear mappings of two feature vectors. After that, a variety of extensions of CCA have been devoted to learning a shared low-dimensional feature space of multiple modalities or views, such as kernel CCA [23, 24], shared kernel information embedding [18, 25]. In addition to CCA, the idea of

*Corresponding author.
Email addresses: iexqyan@zzu.edu.cn (Xiaoqiang Yan), ieshizhehu@gmail.com (Shizhe Hu), ieyqmao@gs.zzu.edu.cn (Yiqiao Mao), ieydye@zzu.edu.cn (Yangdong Ye), hui.yu@port.ac.uk (Hui Yu)



(a) A object is described by text, video, audio

(b) A news is reported by different languages

(c) A product can be represented by multi-view graphs

(d) An image is depicted by heterogonous features

(e) A social image along with user tags

(f) One action can be captured from different camera viewpoints

Figure 1: Examples of multi-view data

MVL has penetrated a variety of learning methods [10, 26, 27, 28], such as dimensionality reduction [13], clustering analysis [29] and ensemble learning [30]. Although these aforementioned methods have achieved promising results, they use hand-crafted features and linear embedding functions, which are not able to capture the nonlinear nature of complex multi-view data. Nonlinearity is a mathematical term describing a situation where there is not a straight-line or direct relationship between an independent variable and a dependent variable. In a nonlinear relationship, changes in the output do not change in direct proportion to changes in any of the inputs. Nonlinearity is a common issue when examining the relationships between the input and output of a learning model. In the domain of machine learning and computer vision, there are various types of nonlinear data, such as text, image, video and audio. With the rapid development of information technology, massive amounts of mult-view data with nonlinear property are generated in real-world applications every day as shown in Figure 1. This linear nature of multi-view data make the learning task on multi-view data remain still challenging.

Recently, due to the powerful feature abstraction ability, deep learning methods [31] have vast inroads into many applications with outstanding performance, such as computer vision [20, 32, 33, 34] and artificial intelligence [19, 35, 36, 22]. Deep learning methods can learn effectively complex, subtle, non-linear and abstract representations of the target data by allowing multiple hierarchical layers. Along with the success of deep learning in many application domains, deep MVL methods also have been increasingly exploited with promising results [35, 19, 32, 36, 33, 34, 20, 22].

In view of the large body of deep learning based methods for MVL proposed in recent years, we attempt to provide a comprehensive review of these works and present our analysis. As shown in Figure 2, we first review the representative MVL methods in the scope of deep learning in this paper, such as multi-view auto-encoder (AE), conventional neural networks (CNN) and deep brief networks (DBN). Then, we investigate the advancements of

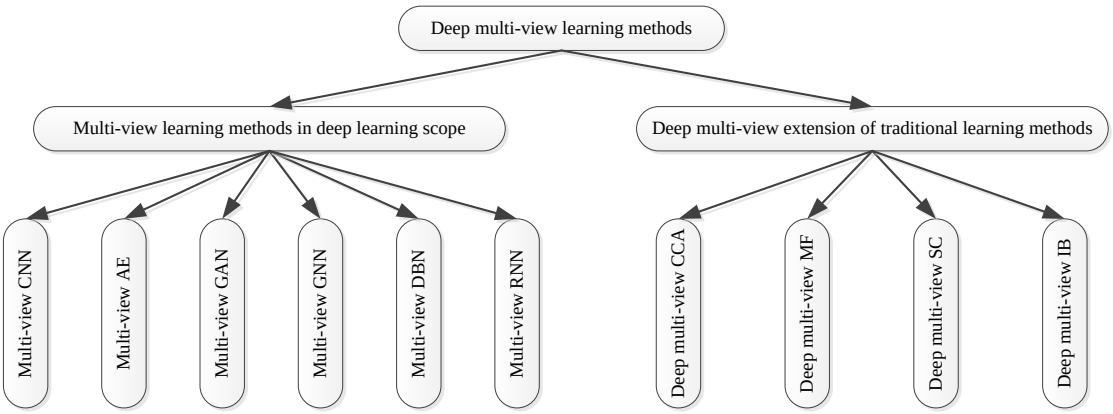


Figure 2: The taxonomy of deep multi-view learning methods.

MVL mechanism when traditional learning methods meet deep learning models, such as deep multi-view canonical correlation analysis (CCA), matrix factorization (MF) and information bottleneck (IB). Finally, we review several important applications, widely-used datasets and open problems of deep MVL methods for further investigation and exploration.

1.1. Comparison with Previous Reviews

Recently, several important related surveys about MVL have been published to summarize the theories, methodologies, taxonomies and applications of the existing MVL approaches [10, 26, 27, 28, 37, 38, 39, 40]. These surveys focus on the problems of specific MVL methods, such as multi-view fusion [10, 27, 28], multi-modal learning [37, 38], multi-view clustering [26] and multi-view representation learning [39, 40].

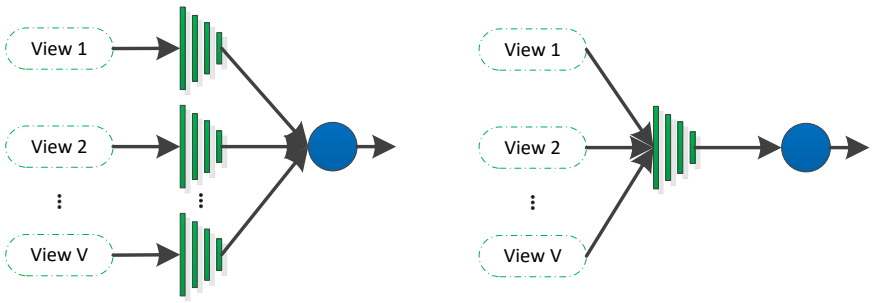
Compared with previous surveys, this paper focuses on reviewing the literature from a cross-perspective of deep learning and MVL, since there are few surveys directly summarizing the deep MVL approaches. In particular, the surveys in [10, 26, 27, 28] focus on MVL approaches in the domain of traditional learning methods. For instance, Baltrusaitis *et al* [38] and Li *et al* [39] summarize the representative shallow learning model for multi-view feature learning approaches, in which the deep learning based MVL methods are ignored or just summarized a small part in their investigations. In contrast, we highlight the deep MVL methods which have gained more attention in recent years. From the viewpoint of deep models, the most related efforts to this paper are [37] and [40]. However, both of them concentrate on the models and applications of multi-modal representation fusion. The main differences between our review and the two aforementioned reviews are summarized as the following two aspects. Firstly, this review concerns more aspects of MVL, while the other two reviews focus on the multi-view feature learning. Secondly, we also review the deep multi-view extension of traditional methods, such as deep multi-view MF, deep multi-view spectral learning and deep multi-view IB, which have never been surveyed.

This paper is organized as follows. In Section 2, the MVL methods in deep learning scope are presented, such as multi-view auto-encoder, auto-encoder and conventional neural networks. In Section 3, the advancements of MVL mechanism are introduced when traditional learning methods meet deep learning models, such as deep multi-view matrix factorization and deep multi-view information bottleneck. In Section 4, we presents several interesting applications of MVL in different domains. In Section 5, several widely-used datasets in the domain of MVL are summarized. In Section 6, the performances of representative deep MVL methods are compared. Finally, several open problems in deep MVL research are provided in Section 6, which we aim to help advance the development of deep MVL.

2. Multi-view Learning Methods in The Deep Learning Scope

2.1. Multi-view Convolutional Neural Network

As a typical deep learning algorithm, convolutional neural network (CNN) [31] aims to learn a high-level feature representation with various parameter optimization [41, 42, 43] and has demonstrated superior performance [44, 45]



(a) One-view-one-network strategy (b) Multi-view-one-network strategy

Figure 3: The two fusion types of multi-view CNNs.

Table 1: The different types of fusion strategy in multi-view CNNs. $*$ means convolution operation while $\otimes$ means matrix accumulation

Fusion types	Mathematical expression	Dimension
Sum	$y^{sum} = x_t^a + x_t^b$	$y^{sum} \in \mathbb{R}^{H\times W\times D}$
Max	$y^{max} = \max(x_t^a, x_t^b)$	$y^{max} \in \mathbb{R}^{H\times W\times D}$
Concatenation	$y^{cat} = cat(3, x_t^a, x_t^b)$	$y^{cat} \in \mathbb{R}^{H\times W\times 2D}$
Convolutional	$y^{conv} = y^{cat} * f + b$	$y^{conv} \in \mathbb{R}^{H\times W\times 2D}$
Biliner	$y^{bil} = \sum_{j=1}^{H}\sum_{i=1}^{W} x_{i,j}^a \otimes x_{i,j}^b$	$y^{bil} \in \mathbb{R}^{D\times D}$

in various domains. Compared with single-view CNN architectures, the multi-view CNN is defined as modelling from multiple feature sets with access to multi-view information of the target data, such as 3D shape recognition [46], multivariate electroencephalography (EEG) [47], multi-feature aggregation [48]. The multi-view CNN architecture aims to integrate multi-view information from different views so as to obtain more discriminative common representations. The existing multi-view CNN architectures are usually partitioned into the following two types: one-view-one-net mechanism and multi-view-one-net mechanism as shown in Figure 3.

The multi-view CNN based on one-view-one-net mechanism adopts one convolutional neural network for each view and extracts feature representation of each view separately, then multiple representations are fused through subsequent part of the network [49, 50, 51, 52, 2, 53, 46, 47, 48]. Taking 3D shape recognition as an example, Yang *et al*. [46] present a multi-view CNN method for comprehensive feature extraction and aggregation of 3-Dimensional (3D) model. As we can see from Figure 4, given a 3D model, it is firstly transformed into N views which generates N images. Then, these images are thrown into N CNN architectures to obtain the feature representations of each view. These view-specific features are integrated and passed into the following feature aggregation model to get a compact, discriminative shape feature. For the one-view-one-network strategy, there are also many other efforts to design multi-view CNN architectures and explore their applications. For instance, Feichtenhofer *et al*. [2] explore several mechanisms of combining CNN features to comprehensively extract the spatio-temporal features of human activities.

Multi-view-one-net mechanism feeds multi-view data into the same network to get the final representation. For example, Dou *et al*. [54] propose a contextual 3D CNNs by combing multi-level information for pulmonary nodule detection. The network includes 3D convolutional layer, 3D max-pooling layer and fully-connected layer to extract the final feature representation hierarchically.

In essence, the difference between one-view-one-net mechanism and multi-view-one-net mechanism lies in the fusion methods of different views. Also taking 3D shape recognition as an example, let $x_t^a \in \mathbb{R}^{H\times W\times D}$ and $x_t^b \in \mathbb{R}^{H\times W\times D}$ denote two input data of the fusion layer, while y indicates the output results of the fusion layer, where H, W, D are the dimension of current layer. The two different fusion types can be summarized as in Table 1.

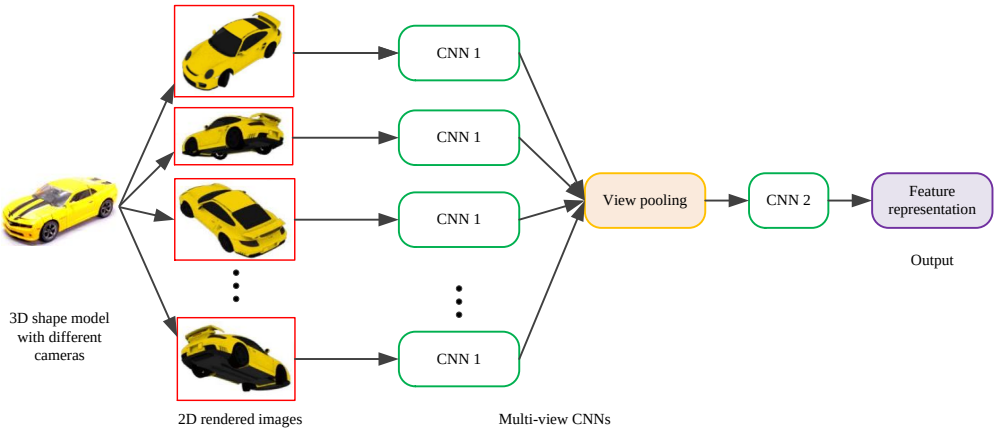


Figure 4: One example of multi-view CNN based on one-view-one-network strategy (adapted from [46]).

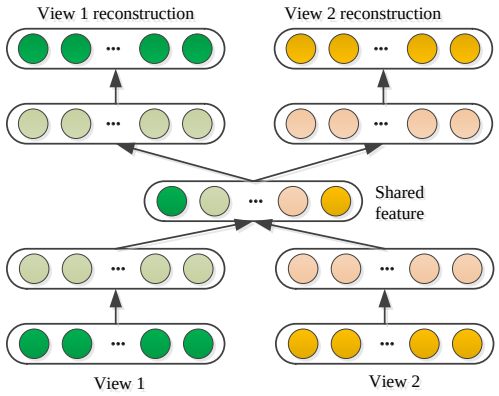


Figure 5: The framework of bimodal auto-encoder (BAE) (adapted from [57]).

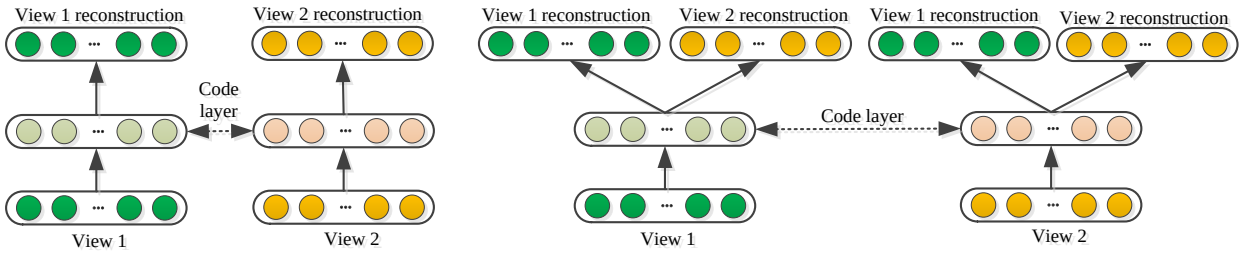
2.2. *Multi-view Auto-encoder*

Auto-encoder (AE) is a variation of neural network and has obtained promising results in virous applications, such as data retrieval [35], human pose recovery [55] and disease analysis [56]. AEs are unsupervised feature learning method in the deep learning literature that consist of two objective functions: encoding function $f(\cdot)$ and decoding function $g(\cdot)$. Specifically, the encoding function aims to map an input data $X \in \mathbb{R}^{D_1}$ to a compressed hidden representation $V \in \mathbb{R}^{D_2}$, $V \approx f(X)$, where D_1 and D_2 are the dimension of original data and its compressed representation. The decoding function $g(\cdot)$ aims to reconstruct the data X from its compressed hidden representation such that $g(V) \approx X$. The hyper-parameters of AE architecture are obtained by minimizing the error $\mathcal{L}(X, g(V))$ of the reconstruction, which can be measured by some losses, like square loss. For example, the cost function of AE is formulated with the square loss as the reconstruction error as follows

$$\sum_{i=1}^{n} \|x_i - g(v_i)\|_F^2 + \sum_{i=1}^{n} \|v_i - f(x_i)\|_F^2 \tag{1}$$

where n is the volume of data instances, $x_i \in X$ and $v_i \in V$ are the input data instances and its optimal representations.

Some recent works are proposed to learn the common representation of multi-view data by using AEs. For example, Ngiam *et al*. [57] devise a novel bimodal auto-encoders (BAE) as shown in Figure 5. The BAE aims to find a reconstruction of both audio and video views by minimizing the reconstruction error of the two input views and



(a) Correspondence auto-encoder (b) Full-modal version of the correspondence auto-encoder

Figure 6: The framework of correspondence auto-encoder (Corr-AE) and its full-modal version (adapted from [35]).

reconstructed representation. Let (X^1, X^2) be the source data including two views, where $X^1 = \{x_i^1\}_{i=1}^n \in \mathbb{R}^{d_1 \times n}$ and $X^2 = \{x_i^2\}_{i=1}^n \in \mathbb{R}^{d_2 \times n}$. The objective function of BAE can be formulated as follows

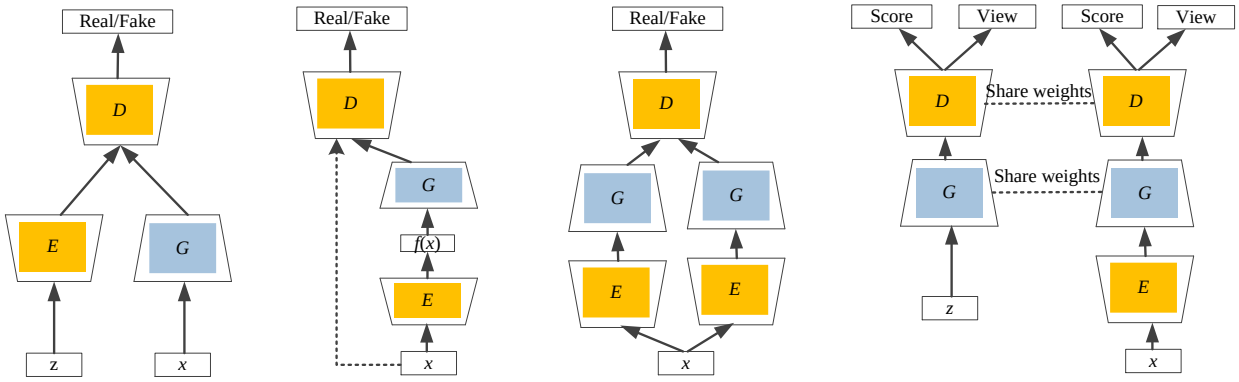
$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \|x_i^1 - f(h(x_i^1, x_i^2; W_h); W_f)\|_F^2 + \|x_i^2 - g(h(x_i^1, x_i^2; W_h); W_g)\|_F^2 \tag{2}$$

where the variables $\theta = \{W_f, W_g, W_h\}$ are the hyper-parameters, in which W_f and W_g denote the weight parameters learning from the encoding function $f(\cdot)$ and denoting function $g(\cdot)$, W_h controls the mapping from the two views into a shared feature space. The BAE is trained in a denoising fashion [58] by utilizing a noisy source data that contains a single view. To allow any combination of modalities to be present or absent at test time, the BAE also needs to train an exponential number of models.

Inspired by BAE, Feng *et al*. [35] present a correspondence auto-encoder (Corr-AE) to conduct cross-modal retrieval, which simultaneously learns the shared information of multiple modalities and the specific information in each individual modalities. The main idea of the Corr-AE is to minimize the correlation learning error between multiple modalities and the feature learning errors of each modality. As shown in Figure 6, the proposed Corr-AE model is composed of the following two subnetworks, each of them is a basic auto-encoder. The two subnetworks are combined by designing a code layer with a predefined similarity measurement. Feng *et al*. [35] also propose a full-modal version of Corr-AE, which can be regarded as an integration of standard auto-encoder and Corr-AE as shown in Figure 6.(b).

Consequently, Zhang *et al*. [59] propose an auto-encoder in auto-encoder network (AE^2-Nets), which focuses on the task of unsupervised representation learning. AE^2-Nets aims to automatically map heterogeneous views into a common representation while adaptively balancing the consistence and complementarity among multiple views. AE^2-Nets adopts inner auto-encoder networks to extract information from each single view, and employs an outer auto-encoder networks to encoder the multi-view information.

Recently, multi-view extensions of AE architectures have been proposed for various tasks, such as 3D reconstruction [60, 19], anomaly detection [3, 61], medical data analysis [56, 32], annotation [62, 63] and human pose recovery [55]. Yang *et al* [60] propose a multi-view 3D reconstruction method with an attentional aggregation model, which fuses multiple deep features encoded from a collection of images recorded by different viewpoints. Bhatt *et al*. [19] propose a set-based deep cross-modal auto-encoders (CMAE), which reconstructs one view of the data under the condition of obtaining the other view, while the interaction between the representations at every intermediate step or each hidden layer is increased. Deepak *et al*. [3] propose to learn deep features for multi-view video anomaly detection by spatiotemporal auto-encoders, which combines handcrafted features and deep features from spatiotemporal auto-encoders with raw input videos. Zhang *et al*. [56] propose a margin-sensitive auto-encoder (MSAE) approach to recognize protein fold and Alzheimer's disease. In MASE, an encoder is utilized to map multi-view physiological signals into a common shared subspace, while a decoder architecture is employed as a restriction of the reconstruction. Besides, MASE uses a self-adjusting scheme to adaptively weight the importance of different views. Wei *et al*. [32] present a multi-view deep neural network (DNN) for magnetic resonance (MR) image segmentation, which uses a convolutional auto-encoder to restore MR slices. Hong *et al* [55] propose a deep auto-encoder architecture for multi-modal human pose recovery, which includes 2D and 3D feature learning parts.



(a) BiGAN (b) DR-GAN (c) TP-GAN (d) CR-GAN

Figure 7: The framework of (a) BiGAN, (b) DR-GAN, (c) TP-GAN and (d) CR-GAN. (Adapted from [67])

2.3. Multi-view Generative Adversarial Networks

As an unsupervised deep learning model, generative adversarial networks (GAN) [64] has been successfully applied into many domains and obtained promising results, such as image-to-image translation [65] and image inpainting [66]. Typically, the basic GAN includes a generative model G and a discriminative model D, thus, it has the most prominent character of adversarial training. The generative model G characterizes the distribution of source data, while the discriminative model D is designed to estimate the probability distribution from the training data.

In this section, we use variable x to indicate the training images and use variable z to refer to a prior noise. In the setting of GAN, we treat the prior noise z as the input of G, while treating the fake image $G(z)$ as the output of G. The process of the generator can be formulated as a function $G(z;\theta_g)$, where θ_g is the parameter of G. Similarly, the input and output of D are x and single scalar $D(x;\theta_d)$, respectively, in which $D(x;\theta_d)$ is the probability that x belongs to the training data. The ultimate goal of GAN is get the following parameter

$$\theta_g^* = \arg\min_{\theta_g}\max_{\theta_g} v(\theta_g,\theta_d) \tag{3}$$

where $v(\theta_g,\theta_d)$ is the loss function as follows

$$v(\theta_g,\theta_d) = E_{x\sim p(x)}[\log D(x)] + E_{z\sim p(z)}\log(1-D(G(z)))] \tag{4}$$

where $p(z)$ and $p(x)$ are the distribution of the given prior noise z and the training data x, respectively.

Generating images with multiple views from a single-view input is a fundamental research topic for broad applications in computer vision, robotics and graphics. And the widely-used GAN has shown impressive results due to its characteristic of of adversarial training [36, 33, 68, 67]. The GAN-based multi-view generation methods first use encoder E to map input images into latent space Z, then decoder G is adopted to generate novel views. However, the single-pathway GAN may learn incomplete representation. To generate complete representation for multi-view generation, great effort has been devoted. Donahue *et al*. [36] propose bidirectional GAN (BiGAN) to learn an inference network E and generator G jointly. In other words, BiGAN provides a means of learning a inverse mapping that projects data back into the latent space. Tran *et al*. [33] propose a DR-GAN method, which also aims to synthesize multi-view images by learning an identity preserved representation. In DR-GAN, the output of its encoder also acts as the input of the decoder, so it cannot deal with new data. After that, Huang *et al*. [68] propose to use two pathway GANs for front view synthesis, in which it the two pathway adopts two distinct encoder-decoder networks, and these two pathways capture global features and local details. Tian *et al*. [67] propose a two-pathway GAN to maintain the completeness of the learned embedding space. The two learning pathways collaborate and compete in a parameter-sharing manner, yielding considerably improved generalization ability to unseen dataset. The differences among these GAN-based multi-view generation are shown in Figure 7.

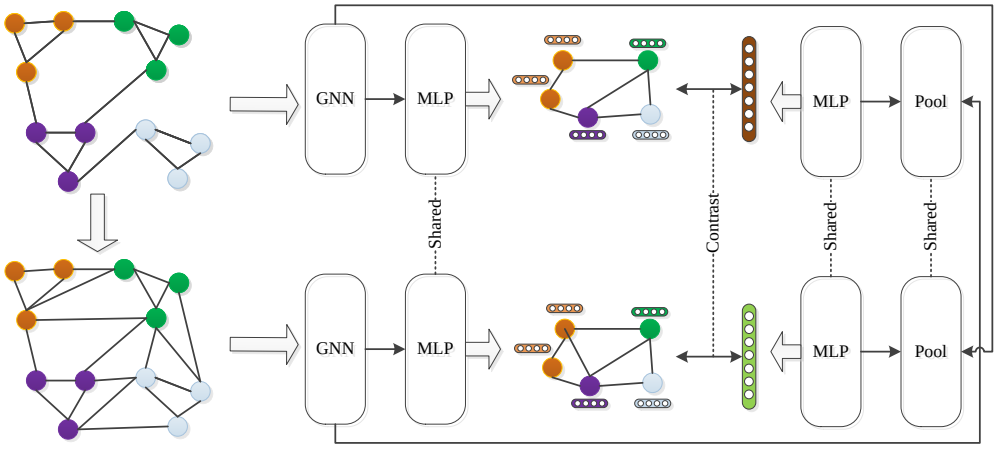


Figure 8: The framework of contrastive multi-view representation learning on graphs

Apart from multi-view generation, GAN also has been applied to many other multi-view applications. For example, Wang *et al*. [69] propose an adversarial correlated auto-encoder (ACAE) to obtain a share feature space for multi-view data, which is applied to cross-view retrieval and classification. Xuan *et al*. [70] focus on multi-view pearl classification, thus propose a multi-view GAN to expand labeled pearl images, which is utilized to train a multistream CNN. Sun *et al*. [71] propose a multi-view embedding network based on GAN, which simultaneously preserves the information from individual network view and accounts for connectivity across different views. Chen *et al*. [72] propose a multi-view extension of BiGAN to conduct density estimation of multi-view source data.

2.4. Multi-view Graph Neural Networks

Graph neural networks (GNN) [73] reconciles the expressive power of graphs in modeling interactions with deep models in terms of learning representation and has gained increasing attention due to its capability of modeling graph structured data. They process variable-size permutation-invariant graphs and learn low-dimensional representations through an iterative process of transferring, transforming and aggregating the representations from topological neighbors. Recently, GNN has achieved outstanding performance in graph-structured data analysis, such as social network [74] and knowledge graphs [75]. First, we briefly review the basic background knowledge of GNN. Let $G = \{V, E\}$ indicate a graph, which is the input data of GNN, the variables $V = \{v_i\}$ and $E = \{e_{ij}\}$ indicate the collection of nodes and edges. Each edge $e_{ij} = (v_i, v_j)$ connects v_i and v_j and each node v_i contains a feature x_i denoting its attribute. The aggregation procedure of GNN can be formulated as follows

$$\begin{aligned} \mathbf{h}^{(0)} &= \mathbf{X} \\ \mathbf{h}^{(l+1)}[i] &= \sigma(\sum_{j \in N_i} \alpha_{ij} \cdot \mathbf{h}^{(l)}[i]\mathbf{W}^{(l)}) \end{aligned} \tag{5}$$

where the variable $\mathbf{X}$ indicates the input features of all nodes in graph G, σ is the non-linear function like Relu, $\mathbf{h}^{(l)}[i]$ indicates the hidden feature of node i in the *l-th* layer, α is a variant of adjacency matrix, $\mathbf{W}^{(l)}$ is the learnable linear transfer matrix.

Recently, GNN also has achieved promising performance in the scenario of MVL, such as multi-graph clustering and multi-view graph conventional networks. We take multi-view representation learning on graphs as an example. Hassani *et al*. [76] present a contrastive multi-view feature learning method on graphs, which indicates that increasing the number of views to more than two cannot improve performance. As shown in Figure 8, the proposed model is performed on both node and graph levels. Firstly, a graph diffusion is adopted to create an additional graph view of the target view, which is fed into two GNNs followed by a shared multi-layer perceptron (MLP) to learn a node representation. Then, the learned feature representations are fed into a graph pooling followed by a shared MLP to learn graph representations.

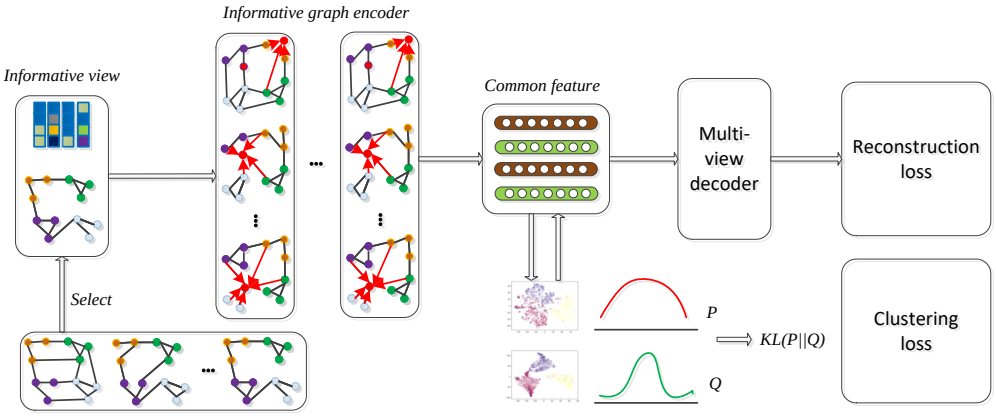


Figure 9: The framework of one2multi graph auto-encoder (adapted from [77]).

Multi-graph clustering based on GNN is an active research direction and has achieved considerable attention recently. For example, Fan *et al*. [77] design a one2multi graph auto-encoder, which is capable of learning node embeddings by using content information to reconstruct the graph structure from multiple views. The proposed model consists of two main parts: a multiple graph auto-encoder and a self-supervised clustering mechanism on the graphs. As shown in Figure 9, one2multi is composed of one graph-based encoder architecture and a multi-view graph-base decoder architecture, in which the most informative view is selected by a heuristic metric modularity.

We take the global poverty analysis, molecular property prediction, multi-view camera re-localization and compression artifacts reduction as the examples to review the applications of multi-view GNN. Specifically, Khan *et al*. [78] propose a convolutional network based on graph structure to analyze global poverty. This approach is applied into three tasks: (1) predicting the adoption of financial inclusion; (2) predicting whether an person is living below the poverty line; (3) predicting the gender of mobile phone subscribers. For molecular property prediction, Ma *et al*. [79] present a multi-view graph neural network from the following observation: both atoms and bonds significantly affect the chemical properties of a molecule, thus it is wise to exploit both node (atom) and edge (bond) information simultaneously to build an expressive model. Xue *et al*. [80] re-design GNN collaborated with CNN in guiding the process of feature extraction and information propagation to obtain the feature representation of multi-view images.

2.5. Multi-view Deep Belief Net

The deep belief net (DBN) is proposed by Geoffrey Hinton et al [81], which adopts the restricted Boltzmann machine (RBM) as its fundamental component. RBM is a simple deep model, which just contains a two layer neural network (a hidden layer and a visible layer). These two layers are combined by the units of the aforementioned hidden and visible layer. In the framework of RBM, the distribution of the two layers is computed by the following energy function

$$E(x,h) = -\sum_j c_j x_j - \sum_i b_i h_i - \sum_j \sum_i h_i w_i x_j \tag{6}$$

where the variable x indicates the visible unit, the variable h indicates the hidden unit, w_{ij} denotes the weight, the variables c_j and b_i are biases in the RBM architecture.

Usually, the DBN architecture is composed of several RBMs by stacking them together as in Figure 10. DBN is a type of generative model, which aims to capture the distribution between the visible objects and their corresponding labels. A DBN architecture with one hidden layers can be formulated as follows

$$P(x,h^1,h^2,\cdots,h^l) = P(x|h^1)P(h^1|h^2)\cdots P(h^{l-2}|h^{l-1})P(h^{l-1},h^l) \tag{7}$$

where x denotes the input data, $P(h^{l-1}|h^l)$ indicates the conditional probability distribution of the *l-th* RBM. $P(h^{l-1},h^l)$ indicates the joint distribution of the RBM in the top layer of DBN.

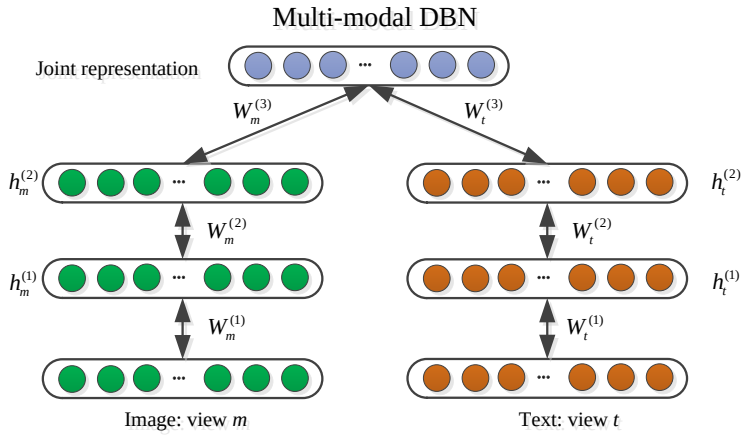


Figure 10: The framework of multi-modal deep brief net (DBN) (adapted from [4]).

To learn the multi-view representations, Srivastava *et al*. [4] present a multi-view DBM based on generative model, which can fit the joint distributions of various data sources, such as audio, image and text. Taking the image and text modality as an example, Srivastava *et al*. [4] first design different DBN models to extract the high-abstract representations of different modalities. Then, a top RBM with one-layer neural network is adopted to characterize the joint representation of multi-view data by passing the combined features of each individual modality.

Recently, some feature learning methods based on multi-view DBN also have been devised. For instance, Amer *et al*. [82] present a hybrid model for detecting the sequential event, in which the inter-modality and cross-modality features are extracted by a conditional RBM with discrimination label information. Waisy *et al*. [83] present a multi-modal fact recognition method, in which a DBN-based model is utilized to capture the complementary effect of local and deep feature representations. Syafiandini *et al*. [84] propose a multi-modal deep Boltzmann machine (DBM) to recognize the importance of different genes, in which the gene expression data and several patient phenotypes are processed simultaneously. Zhang *et al*. [85] propose a multi-modal correlation DBN, which uses a RBM to model a separate view.

2.6. Multi-view Recurrent Neural Network

For dealing with time series data, Sutskever *et al* propose a recurrent neural network (RNN) [86], which has been successfully applied into various analysis tasks on time series data. The RNN model consists of three kinds of layers in different time frames, which are the input word layer $\mathbf{w}$, the recurrent layer $\mathbf{r}$ and the output layer $\mathbf{y}$. We use $\mathbf{w}(t)$, $\mathbf{r}(t)$ and $\mathbf{y}(t)$, respectively to indicate the activation of these types of layers at time t. $\mathbf{w}(t)$ denotes the current word vector, which is a one-hot representation and can be coded in a simple 1-of-N representation. As in [86], $\mathbf{y}(t)$ can be formulated as follows

$$\mathbf{x}(t) = [\mathbf{w}(t), \mathbf{r}(t-1)]; \mathbf{r}(t) = f_1(\mathbf{U} \cdot \mathbf{x}(t); \mathbf{y}(t) = g_1(\mathbf{V} \cdot \mathbf{r}(t); \tag{8}$$

where the variables $\mathbf{U}$ and $\mathbf{V}$ indicate the weights that will be learned by the RNN architecture, the variable $\mathbf{x}(t)$ denotes a vector that integrates $\mathbf{r}(t-1)$ and $\mathbf{w}(t)$, while the functions $f_1(\cdot)$ and $f_2(\cdot)$ indicate the sigmoid and softmax function.

Mao *et al*. [5] propose a multi-view RNN architecture, which aims to generate captions for visual images. As shown in Figure 11, this multi-view RNN includes the following three subnetworks: a vision network, a language network and a multi-view network. Specifically, the vision network usually adopts a deep CNN, such as Resnet, Alexnet and Inception, which aims to map the visual information of an image into its deep feature representation. The language network is to capture the task-specific representation and the temporal dependency. In the multi-view network, a hidden network is always adopted to find the relationship between the vision representation and the learned language. Following [5], Karpathy *et al*. [6] present a multi-view alignment model based on RNN to bridge the inter-view relationship between visual and textual data.

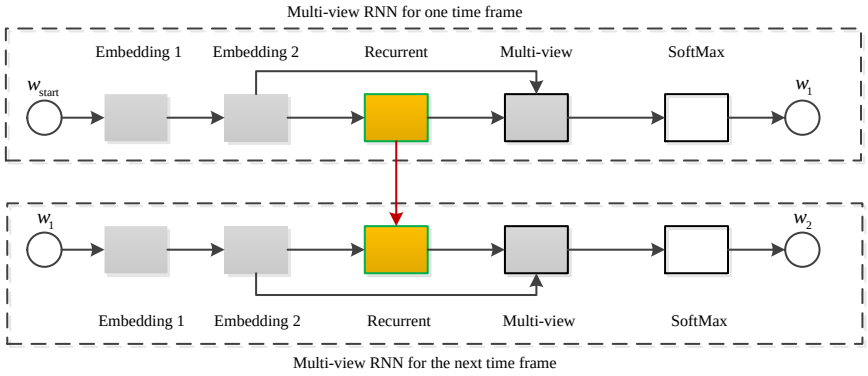


Figure 11: The framework of multi-view recurrent neural network (RNN) (adapted from [5]).

Recently, several RNN-based deep MVL methods are proposed. For instance, Abdulnabi *et al*. [87] design a multi-view RNN to recognize the categories in the indoor scenes, where the standard RNN architecture is utilized to extract the cross-view and intra-view feature. Sano *et al*. [88] propose a multi-view long short-term memory named BiLSTM to detect ambulatory sleep, where the proposed BiLSTM is adopted to characterize the data captured by wearable devices. Narayanan *et al*. [89] propose a gate RNN framework to recognize the driver behaviors by analyzing the multi-view sensor data.

Actually, there are many problems that make MVL challenging, such as incomplete views, low-quality input data and the presence of view disagreement. Thanks to the strong ability of data abstraction, the deep learning based MVL methods have obtained promising performance in various tasks. However, apart the issues that all the multi-view learning methods have to tackle with, there are also several limitations of the deep learning based MVL. First, the deep MVL methods need much more training data [90]. Obviously, the volume of the current multi-view training datasets is limited because of the high cost and labor-consuming of manual labeling. Second, although existing deep MVL models have shown superior advantages in various applications, they fail to provide an explanation for the decision of different models. Thus, the deep models without explanation are not easily applied to critical domains, such as military system or medical treatment [8]. Finally, despite their great success, there is still no comprehensive theoretical understanding of the learning mechanism with deep neural networks (DNNs) or their inner organization [91].

3. Deep Multi-view Extensions of Traditional Learning Methods

Although the MVL in view of deep learning technique has obtained promising performance in various domain, they fail to provide a explainable mechanism in the learning process, which results in the sensitivity to critical domains. In this section, we review the multi-view extension works of conventional learning methods. In general, the conventional learning methods usually have complete theoretical foundation, such as canonical correlation analysis, spectral clustering and information bottleneck. The conventional learning methods can obtain significant performance improvement by incorporating the strong ability of data abstraction, which also attracts lots of attention in the domain of machine learning and computer vision.

3.1. Deep Multi-view Canonical Correlation Analysis

In this part, the typical canonical correlation analysis (CCA) [11] method is firstly presented and then several representative deep multi-view extensions of CCA are reviewed.

3.1.1. Canonical Correlation Analysis

CCA is a well-known and widely used approach to capture the correlations between two given variables by mapping them into a common subspace. Given two variables, CCA intends to discover the linear mapping between them to a common space in which their correlation are maximally preserved. It is precisely because of this characteristic of

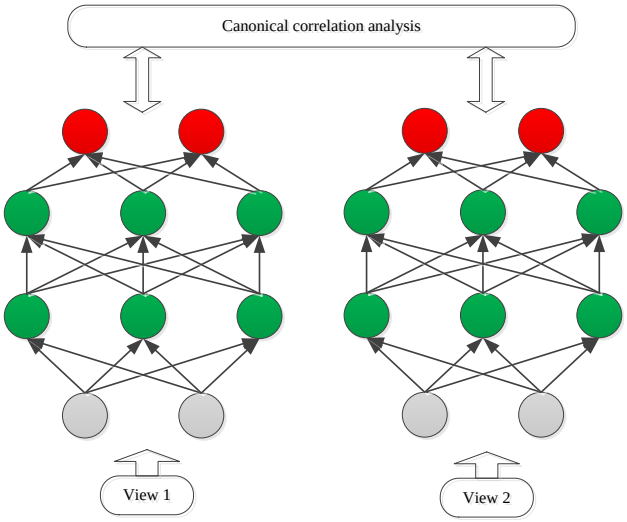


Figure 12: The framework of deep canonical correlation analysis (DCCA)

CCA that it usually is utilized to conduct the task of MVL and has obtained promising performance in various tasks including dimension reduction [12, 13, 14], classification [92, 93] and clustering [94, 95], when multiple views are available.

Let $X_1 \in \mathbb{R}^{d_1 \times n}$ and $X_2 \in \mathbb{R}^{d_2 \times n}$ indicate two random variables with n samples, d_1 and d_2 are the dimension of X_1 and X_2, respectively. The objective of CCA is to discover the linear mappings of the two different views that maximize their relatedness. The objective of CCA can be written as follows

$$\begin{aligned}(w_1^*, w_2^*) &= \arg\max_{w_1, w_2} corr(w_1^T X_1, w_2^T X_2) \\ &= \arg\max_{w_1, w_2} \frac{w_1^T C_{12} w_2}{\sqrt{(w_1 C_{11} w_1)(w_2 C_{22} w_2)}}\end{aligned} \tag{9}$$

where C_{11} and C_{22} are covariances of each view X_1 and X_2, respectively. C_{12} is the cross-covariance between X_1 and X_2.

The traditional CCA can only discover the linear relationship with two variables, thus it is just able to discover the linear correlation of the two views of multi-view data. To tackle this problem, some nonlinear extensions of CCA are proposed, such as Kernel CCA [23, 24, 18, 25], locality preserving CCA [96, 97] and deep CCA [20]. Next, we review several representative deep CCA and deep multi-view extensions of CCA.

3.1.2. Deep Canonical Correlation Analysis

To discover the high-level relevances among different samples recorded or represented from multiple viewpoints, several attempts have been devoted for decades. Becker and Hinton *et al*. [98] propose a multi-layer extension of CCA by maximizing the canonical correlation between two neural networks. Becker *et al*. [99] also investigate to extract high order features from multiple views by maximally preserving the mutual information between multiple neural networks. Hsieh *et al* [100] use three feedforward networks to formulate a nonlinear CCA. Although several neural networks based on CCA have been investigated before, the deep learning version of traditional CCA, named deep CCA, is developed by Andrew *et al* [20]. Specifically, the deep CCA tries to capture non-linear correlations between different views by combining DNNs and CCA.

Let f and g represent two independent neural networks, deep CCA aims to optimize parameters θ_f and θ_g of networks f and g so as to maximally preserve the canonical correlation between the output of f and g, as shown in Figure 12. The weights, θ_f and θ_g, of these two networks are trained via a standard back-propagation method to

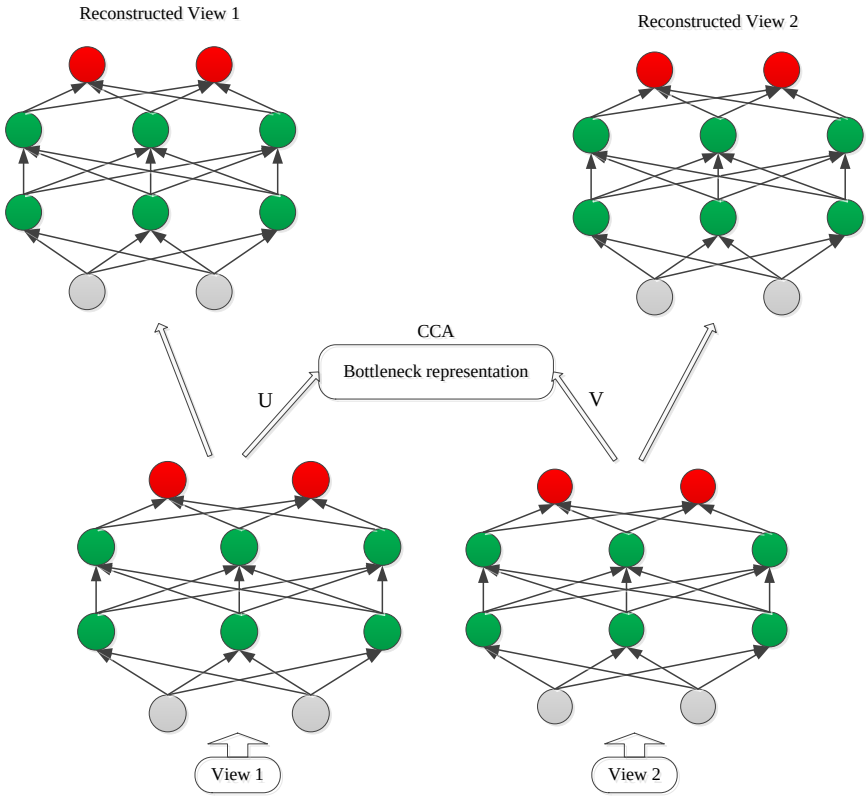


Figure 13: The framework of deep canonical correlated auto-encoders (DCCAE) (adapted from [22]).

maximize the following objective function

$$(\theta_f^*, \theta_g^*) = \arg\max_{\theta_f, \theta_g} corr(f(X_1; \theta_f), g(X_2; \theta_g)) \tag{10}$$

Recently, the performance of deep CCA based approaches have been verified in various feature learning applications, such as image and text matching [101, 102], multiple language embedding [103] and cross-modal subspace clustering [104, 105].

3.1.3. Deep Canonical Correlated Auto-encoders

On the basis of the framework of the deep CCA and reconstruction-based approaches, Wang *et al.* [22] propose a deep canonical correlated auto-encoders (DCCAE), which is an extension of deep CCA. Specifically, the proposed DCCAE is composed of two auto-encoders (as shown in Figure 11), in which the canonical relationship of the compressed features and reconstruction errors of the auto-encoders are optimized simultaneously. The objective function

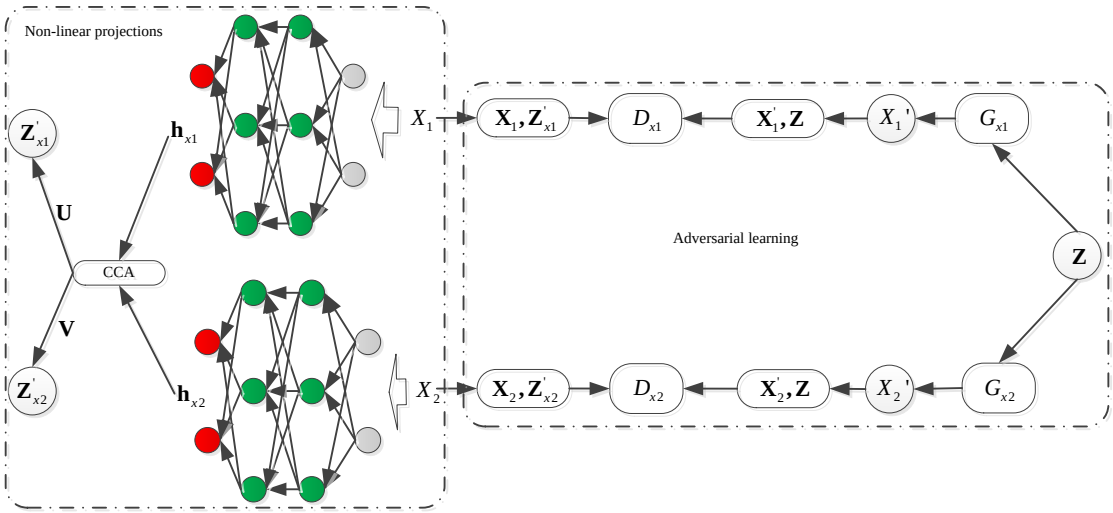


Figure 14: The framework of deep adversarial canonical correlation analysis (DACCA) (adapted from [7]).

of DCCAE is formulated as follows

$$
\begin{aligned}
&\mathcal{L}_{DCCAE} = \max_{\theta_f,\theta_g,\theta_p,\theta_q,U,V} -\frac{1}{n} tr(U^T f(X_1) g(X_2)^T V) \\
&+ \frac{\lambda}{n} \sum_{i=1}^{n} (\|x_1^i - p(f(x_1^i))\|^2 + \|x_2^i - p(f(x_2^i))\|^2) \\
&s.t., U^T (\frac{1}{n} f(X_1) f(X_1)^T + r_{X_1} I) = I \\
&V^T (\frac{1}{n} f(X_2) f(X_2)^T + r_{X_2} I) = I \\
&u_i^T f(X_1) g(X_2)^T v_j = 0, \forall ij
\end{aligned}
\tag{11}
$$

where the variable I denotes the identity matrix, $U = \{u_1, u_2, \cdots, u_L\}$ and $V = \{v_1, v_2, \cdots, v_L\}$ are the CCA directions that map the features of auto-encoders to a compressed space with L units, (r_{X_1}, r_{X_2}) are the regularization terms.

The DCCAE objective provides a balance between the information hidden in each individual view and the correlations across multiple views. From this viewpoint, the idea of the DCCAE is in line with the theory of the well-known information bottleneck (IB) approach [106]. And indeed, the IB approach also aims to extract a subspaces as CCA in the case of Gaussian variables [107].

3.1.4. Deep Adversarial Canonical Correlation Analysis

Due to the promising performance of adversarial learning mechanism, Fan *et al.* [7] present a deep adversarial extension of the traditional CCA, named DACCA. The DACCA can simultaneously synthesize multi-view data instances and capture the discriminative feature of the multi-view data. In the non-linear mappings of feature learning, the objective of DACCA is similar with DCCAE as in Equation 11. We briefly present the generator model and discriminator model in the following.

Generator model. In the part of generator model, it aims to synthesize paired data instances as realistic as possible. In the scenario of CCA, there are two generator models for these two views X_1 and X_2. Let $G_{x_1}(\mathbf{Z})$ and $G_{x_2}(\mathbf{Z})$ denote the generators to get fake sample X_1' and X_2', respectively, where $\mathbf{Z}$ is latent vector $\mathbf{Z} \in \mathbb{R}$. Both of the

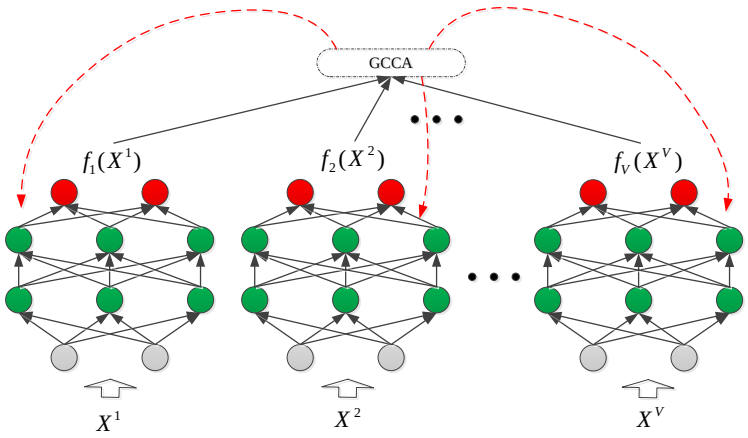


Figure 15: The framework of deep generalized canonical correlation analysis (DGCCA) (adapted from [108]).

generators can be formulated as follows

$$\begin{aligned} X_1' &= G_{x_1}(\mathbf{Z}) = g^p_{\theta_{x_1}}(g^{p-1}_{\theta_{x_1}}(\cdots g^2_{\theta_{x_1}}(g^1_{\theta_{x_1}}(\mathbf{Z})))) \\ X_2' &= G_{x_2}(\mathbf{Z}) = g^p_{\theta_{x_2}}(g^{q-1}_{\theta_{x_2}}(\cdots g^2_{\theta_{x_2}}(g^1_{\theta_{x_2}}(\mathbf{Z})))) \end{aligned} \tag{12}$$

where the variables p and q denote the layer number in $G_{x_1}(\mathbf{Z})$ and $G_{x_2}(\mathbf{Z})$. $g^i_{\theta_{x_1}}$ and $g^i_{\theta_{x_2}}$ are the *i-th* layers of $G_{x_1}(\mathbf{Z})$ and $G_{x_2}(\mathbf{Z})$.

Discriminative model. Different from the original GAN, Fan *et al.* [7] introduce a disciminator D_x to distinguish joint pairs $(X_1, \mathbf{Z}'_{x_1})$ and $(X_1', \mathbf{Z})$, where $\mathbf{Z}'_{x_1}$ is the output of the non-linear projection component. Then, the discriminative value on variable X_1 is given as in [7]

$$\begin{aligned} &\min_{G_{x_1}, E_{x_1}, \mathbf{U}} \max_{D_x} L_{AdvX} \\ &\mathbb{E}_{x_1 \sim P(x_1), \mathbf{h}_{x_1} \sim P_{E_{x_1}}(\mathbf{h}_{x_1}|x_1)}[\log D_x(X_1, U^T \mathbf{h}_{x_1})] \\ &\mathbb{E}_{\mathbf{Z} \sim P(\mathbf{Z}), x_1' \sim P_{G_{x_1}}(x_1'|\mathbf{Z})}[1 - \log D_x(X_1', \mathbf{Z})] \end{aligned} \tag{13}$$

where $\mathbf{h}_{x_1} = E_{x_1}(x_1)$ and $\mathbf{h}_{x_2} = E_{x_2}(x_2)$ are the encoders for views X_1 and X_2. Likewise, the discriminative value on variable X_2 can be obtained. The overall architecture of DACCA is presented in Figure 14.

3.1.5. Deep Generalized Canonical Correlation Analysis

In the aforementioned deep multi-view CCA method, they can just deal with two views of the multi-view data. Actually, MVL approaches should be able to model nonlinear relationships between more views than two. Aiming at this issue, Adrian *et al.* [108] present a deep generalized CCA (DGCCA) method, which is able to discover the nonlinear transformations of multiple data views.

DGCCA is an extension of generalized CCA (GCCA), which is a shallow model that addresses the limitation on the number of views. GCCA aims to find a share representation Z of V different views. Let variable n denote the volume of data samples, d_i indicate the dimension of the *i-th* view, while the variable r denote the dimension of Z. Thus, the objective function of GCCA is formulated as

$$\mathcal{L}_{GCCA} = \min_{U_i \in \mathbb{R}^{d_i \times r}, Z \in \mathbb{R}^{r \times n}} \sum_{i=1}^{(V)} \|Z - U_i^T X_i\|_F^2, \; subject\ to\ ZZ^T = I \tag{14}$$

Similar with GCCA, Adrian *et al*. [108] formulate DGCCA as follows

$$\mathcal{L}_{DGCCA} = \min_{U_i \in \mathbb{R}^{d_i \times r}, Z \in \mathbb{R}^{r \times n}} \sum_{i=1}^{(V)} \|Z - U_i^T f(X_i)\|_F^2, subject\ to\ ZZ^T = I \tag{15}$$

As shown in Figure 15, the objective of the proposed DGCCA can be formulated as: finding weight matrices defining functions f_i, and linear transformations U_i, which is the output of the *i-th* network.

3.2. Deep Multi-view Matrix Factorization

In this part, the traditional single-view, multi-view MF and multi-layer MF are firstly presented. Then several representative deep multi-view extensions of MF are reviewed.

3.2.1. Matrix Factorization

Matrix factorization (MF) [109] methods have been extensively applied to pattern recognition, computer vision, data mining and machine learning, which is also called matrix decomposition. In the various variants of MF, nonnegative matrix factorization (NMF) [110], a specific form of matrix factorization, has received significant attention in various domains [111, 112, 113]. For clarity, we first briefly present some preliminaries about NMF. Suppose there is a matrix $\mathbf{X} = \{x_1, x_2, \cdots, x_n\} \in \mathbb{R}^{d \times n}$ denoting source data, where n indicates the volume of data instances and d denotes the size of feature dimension, the objective of NMF is to find two matrices $\mathbf{Z} = \mathbb{R}^{d \times k}$ and $\mathbf{H} = \mathbb{R}^{k \times n}$ with nonnegative value such that $\mathbf{X} \approx \mathbf{ZH}$. To this end, the objective function of traditional NMF can be formulated as follows

$$\mathcal{L}_{NMF} = \sum_{i=1}^{n} \sum_{j=1}^{m} [\mathbf{X}_{ij} - (\mathbf{ZH}^T)_{ij}] = \|\mathbf{X} - \mathbf{ZH}^T\|_F^2 s.t. \mathbf{Z} \geq 0, \mathbf{H} \geq 0 \tag{16}$$

where the variable $\mathbf{X}_{ij}$ denotes the data sample (i, j)-*th* in $\mathbf{X}$ and $\|\cdot\|_F^2$ is the Frobenius norm.

Due to its intuitive parts-based interpretation, NMF has drawn significant attention in the scenario of MVL. The traditional MF-based MVL approaches can be partitioned into two types: multiple kernel/graph [114, 115, 116, 117] and subspace learning methods [118, 119]. Multiple kernel/graph learning based approaches aim to integrate different kernels/graphs into a unified one. For example, Cai *et al*. [114] learn a shared graph Laplacian matrix, in which a nonnegative constraint is imposed. Huang *et al*. [115] and Nie *et al*. [116, 117] assign weights for the given views, in which unreliable views are allocated lower weight while reliable ones are assigned higher weights. Subspace learning based approaches [118, 119] intend to map multi-view data into a common subspace. For instance, Gao *et al*. [118] present a joint NMF (JNMF) method to encourage unsupervised learning of each view towards a consensus. However, the aforementioned NMF based multi-view learning methods use hand-crafted features and linear embedding functions, which are not able to analyze datasets with complex data distributions, such as appearance, shape and color factors in image object detection. Aiming at this problem, Trigeorgis*et al*. [120] propose a deep semi-NMF to capture the deep features by decomposing the source data $\mathbf{X}$ into m layers. The deep layer matrix decomposition can be formulated as

$$\begin{aligned} \mathbf{X} &\approx \mathbf{Z}_1 \mathbf{H}_1^T \\ \mathbf{X} &\approx \mathbf{Z}_1 \mathbf{Z}_2 \mathbf{H}_2^T \\ &\vdots \\ \mathbf{X} &\approx \mathbf{Z}_1 \mathbf{Z}_2 \cdots \mathbf{Z}_r \mathbf{H}_r^T \end{aligned} \tag{17}$$

where $\mathbf{Z}_i$ and $\mathbf{H}_i$ are the *i-th* layer in the basis matrix and representation matrix, respectively. However, the deep semi-NMF method is a single-view deep extension of NMF. Next, we review several representative deep multi-view extension of NMF based on the deep semi-NMF method.

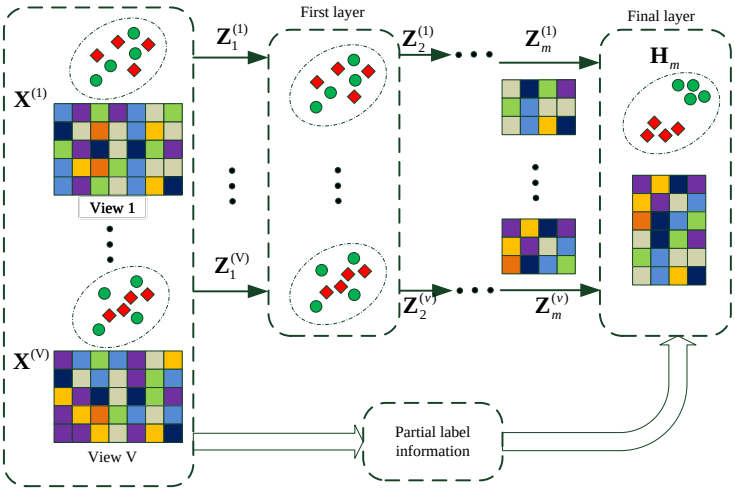


Figure 16: The framework of deep multi-view concept learning method (DMCL) (adapted from [122]).

3.2.2. *Multi-view Clustering via Deep Matrix Factorization*

Inspired by the work in [120], Zhao *et al*. [121] extend the deep matrix factorization technique to a multi-view clustering (MVC) method. In that study, the semi-NMF is employed to discover the deep features of the source multi-view data.

Suppose there is a source data $\mathbf{X} = \{X^{(1)}, X^{(2)}, \cdots, X^{(V)}\}$ with multiple views, where V denotes the amount of views. For each individual view, we assume $X^{(i)} \in \mathbb{R}^{d_i \times n}$, where n denotes the volume of data samples, d_i denotes the dimension of feature representation of the *i-th* view. Then, the objective of MVC via deep MF can be formulated as

$$\begin{aligned} \mathcal{L}_{MVC-DNMF} = \min_{\mathbf{Z}_i^{(v)}, \mathbf{V}_m} \sum_{(v=1)}^{V} (\alpha^{(v)})(\|X^{(v)} - Z_1^{(v)} Z_2^{(v)} \cdots Z_m^{(v)} H_m^{(v)}\|_F^2) \\ + \beta tr(H_m L^{(v)} H_m^T), s.t. H_i^{(v)} \geq 0, H_m \geq 0, \sum_{v=1}^{V} \alpha^{(v)} = 1 \end{aligned} \tag{18}$$

where the variable $Z_i^{(v)}, i = \{1, 2, \cdots, m\}$ indicates the *i-th* layer of the *v-th* view, H_m denotes the combination latent representation for all views and $\alpha^{(v)}$ controls the weights of different views. Although it is a deep multi-view factorization approach, it neither considers ground-truth information of data nor does it explicitly model consistent and complementary information.

3.2.3. *Deep Multi-view Concept Learning Method*

To explicitly deal with the problem of discern of complementary relationship hidden in source multi-view data, Xu *et al*. [122] introduce the thought of concept learning into the multi-view learning based on deep MF and propose a deep multi-view concept learning (DMCL) method as in Figure 16. Specifically, DMCL factorizes the source multi-view data iteratively to extract the high-level feature, in which partial label information is utilized to learn semantic structures and structured sparseness constraint. Actually, DMCL model is an extension of traditional multi-view concept learning (MCL) [123], which explicitly discerns consistent and complementary information in multi-view data to generate conceptual representations. In MCL, the data matrix of each view is separated into two parts: labeled and unlabeled ones, i.e., $\mathbf{X}^{(v)} = [\mathbf{X}^{(v),l} \mathbf{X}^{(v),u}]$. Correspondingly, the encoding matrix can be expressed as $\mathbf{H} = [\mathbf{H}^l \mathbf{H}^u]$.

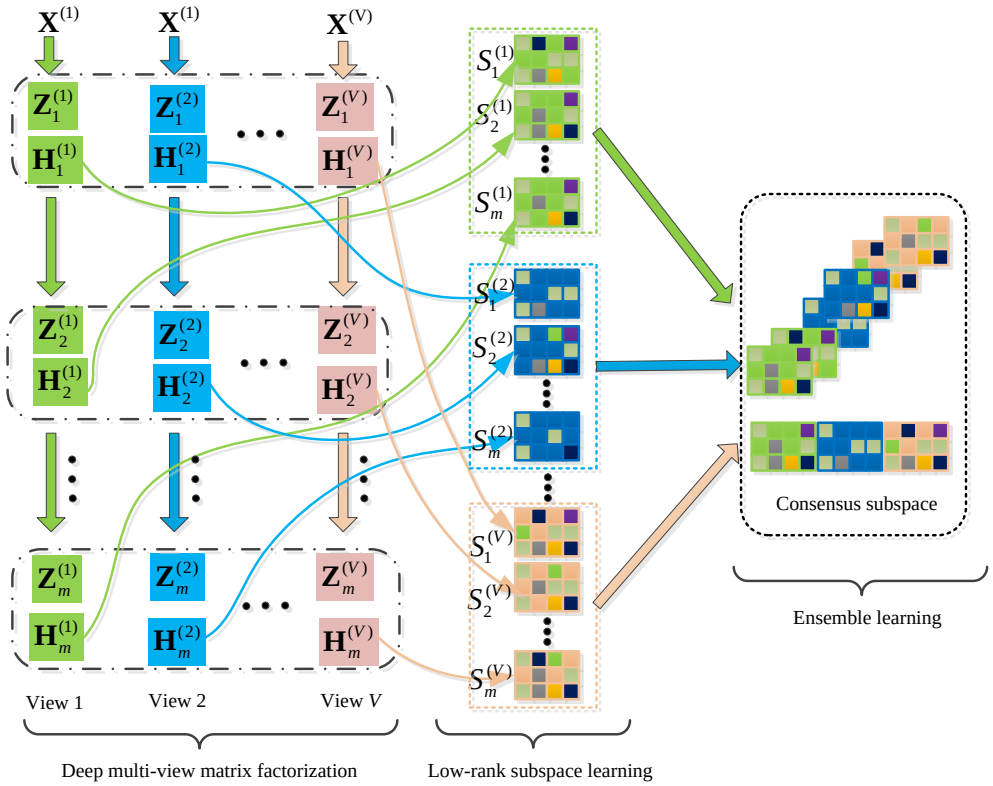


Figure 17: The framework of deep Low-Rank Subspace Ensemble (DLRSE) method for multi-view clustering (adapted from [17]).

As in [123], the objective function of MCL is formulated as

$$\begin{aligned}\mathcal{L}_{MCL} &= \min_{\mathbf{Z}_i^{(v)},\mathbf{H}} \frac{1}{2}\sum_{(v=1)}^{V} \|\mathbf{X}^{(v)} - \mathbf{Z}^{(v)}\mathbf{H}^{(v)}\|_F^2) + \alpha \sum_{i=1}^{V} \|\mathbf{Z}^{(v)}\|_{1,\infty} \\ &+ \frac{\beta}{2}\{tr(\mathbf{H}^l\mathbf{L}^{(a)}(\mathbf{H}^l)^T) - \{tr(\mathbf{H}^l\mathbf{L}^{(p)}(\mathbf{H}^l)^T)\} + \gamma\|\mathbf{H}\|_{1,1}, s.t. U_{ik}^{(v)} \geq 0, 1 \geq Z_{kj} \geq 0, \forall i, j, k, v.\end{aligned} \tag{19}$$

Based on the formulation 19 of MCL, deep multi-view concept learning (DMCL) decomposes each of data matrices iteratively to obtain the deep representation according to $\mathbf{X}^{(v)} \approx \mathbf{Z}_1^{(v)}\mathbf{Z}_2^{(v)}\cdots\mathbf{Z}_m^{(v)}\mathbf{H}_m$, where $\mathbf{Z}_1^{(v)}, \mathbf{Z}_2^{(v)}, \cdots, \mathbf{Z}_r$ denote basis matrices and $\mathbf{H}_r$ indicates the final representation. Thus, the optimization of DMCL can be formulated as

$$\begin{aligned}\mathcal{L}_{DMCL} &= \min_{\mathbf{Z}_i^{(v)},\mathbf{H}} \frac{1}{2}\sum_{(v=1)}^{V} \|\mathbf{X}^{(v)} - \mathbf{Z}_1^{(v)}\mathbf{Z}_2^{(v)}\cdots\mathbf{Z}_m^{(v)}\mathbf{H}_m^{(v)}\|_F^2) + \\ &\frac{\beta}{2}\{tr(\mathbf{H}_m^l\mathbf{L}^{(a)}(\mathbf{H}_m^l)^T) - \{tr(\mathbf{H}_m^l\mathbf{L}^{(p)}(\mathbf{H}_m^l)^T)\} + \\ &\alpha\sum_{i=1}^{V}\|\mathbf{Z}_m^{(v)}\|_{1,\infty} + \gamma\|\mathbf{H}_m\|_{1,1}, s.t. (U_m^{(v)})_{ik} \geq 0, 1 \geq (H_m)_{kj} \geq 0, \forall i, j, k, v.\end{aligned} \tag{20}$$

where m denotes the number of layers. Compared with the MVC via deep MF in [121], DMCL incorporates the partial label information learn semantic structures and structured sparseness constraint, which can explore the complementary and consistent information at the higher level in the setting of MVL.

3.2.4. Deep Low-rank Subspace Ensemble

The existing deep multi-view MF models mainly utilize the representation of the final layer and overlook the latent correlation in the middle layers of the network, which is always valuable for the learning accuracy. Aiming at this issue in the deep multi-view MF, Xue *et al*. [17] extend the deep multi-view MF to a deep low-rank subspace ensemble (DLRSE) approach as in Figure 17. First, for the data in each view, the deep semi-NMF approach is performed to obtain the multi-layer representations of each view as

$$\begin{aligned} &\mathcal{L}_{semi-NMF}(Z_i^{(v)}, H_i^{(v)}) = \sum_{(v=1)}^{V} \|X^{(v)} - Z_1^{(v)} Z_2^{(v)} \cdots Z_m^{(v)} H_m^{(v)}\|_F^2 \\ &s.t. H_i^{(v)} \geq 0, H_m \geq 0, i \in \{1, 2, \cdots, m\}, v \in \{1, 2, \cdots, V\}, \end{aligned} \tag{21}$$

By performing the deep semi-NMF on each individual view, the latent information in each individual view can be captured in the coefficient matrices $\{H_i^{(m)}\}_{i=1}^m$ of each layer. To learn the correlation information of the data instances from multiple views, DLRSE adopts low-rank representation (LRR) method [124] to enhance the pattern structures by learning low-rank subspace of different views. Then, the objective of DLRSE can be formulated as

$$\begin{aligned} &\mathcal{L}_{DLRSE}(F, S_i^{(v)}, \alpha) = \sum_{v=1}^{V} \sum_{i=1)}^{m} \alpha_i^{(v)} \|F - F S_i^{(v)}\|_F^2 + \lambda \|\alpha\|_G \\ &s.t. \alpha \geq 0, 1^T \alpha = 1, F F^T = I, \end{aligned} \tag{22}$$

where $\alpha \in \mathbb{R}^{mV \times 1}$ is the coefficient vector, the variable $S_i^{(v)} \in \mathbb{R}^{n \times n}$ indicates the subspace in the *i-th* layer of the *v-th* view. F is the low-dimensional subspace of the source multi-view data. The optimization details is detailed in [17].

Apart from the aforementioned models, there are also several deep multi-view matrix factorization methods, such as deep collective matrix factorization [125], auto-weighted deep multi-view matrix factorization [126], multi-view multiple clusterings via deep matrix factorization [127]. These methods explore the effectiveness of different mechanisms for the deep multi-view matrix factorization and obtain promising results in virous domains.

3.3. Deep Multi-view Spectral Learning Network

In this part, the traditional single-view and multi-view spectral clustering (SC) are first presented briefly. Then several representative deep multi-view extensions of spectral clustering are reviewed.

3.3.1. Spectral Learning

In the domain of unsupervised clustering, spectral clustering (SC) [128] has obtained promising performance on arbitrary shaped clusters. Given a dataset $X = \{x_1, x_2, \cdots, x_n\} \in \mathbb{R}^{d \times n}$, the objective of SC is to partition all instances into c categories. Specifically, SC first generates an affinity matrix or graph $\mathbf{W}$ for all data instances as follows

$$W_{ij} = \begin{cases} \exp - \dfrac{\|x_i - x_j\|_2^2}{2\sigma^2}, x_i,\ x_j \ are\ connected. \\ 0,\ otherwise. \end{cases} \tag{23}$$

where the variable $W_{ij} \in \mathbf{W}$ denotes the connection weight of the *i-th* and *j-th* data point. Once obtaining $\mathbf{W}$, the SC can be formulated as

$$\mathcal{L}_{SC} = \arg\min_{\mathbf{Z}} Tr(\mathbf{Z}^T \mathbf{L} \mathbf{Z}), s.t., \mathbf{Z}^T \mathbf{Z} = \mathbf{I} \tag{24}$$

where $\mathbf{Z}$ is the feature representation of source data. The clustering assignment can be obtained by conducting k-means on it. $\mathbf{D}$ is a diagonal matrix $D_{ii} = \sum_j W_{ij}$, $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is Laplacian matrix.

Recently, lots of traditional multi-view spectral clustering approaches also have been applied into many practical fields, such as co-regularized SC (CRSC) [129], co-training SC (CTSC) [16], convex sparse SC [130], multi-view SC co-clustering [131] and [132]. Li *et al*. [133] extend it to multi-view scenario as follows

$$\mathcal{L}_{MVSC} = \arg\min_{\mathbf{Z}, \alpha^{(v)}} \sum_{v=1}^{V} (\alpha^{(v)})^r Tr(\mathbf{Z}^T \mathbf{L}^{(v)} \mathbf{Z}), s.t., \mathbf{Z}^T \mathbf{Z} = \mathbf{I}, \sum_{v=1}^{V} \alpha^{(v)} = 1, \alpha^{(v)} > 0 \tag{25}$$

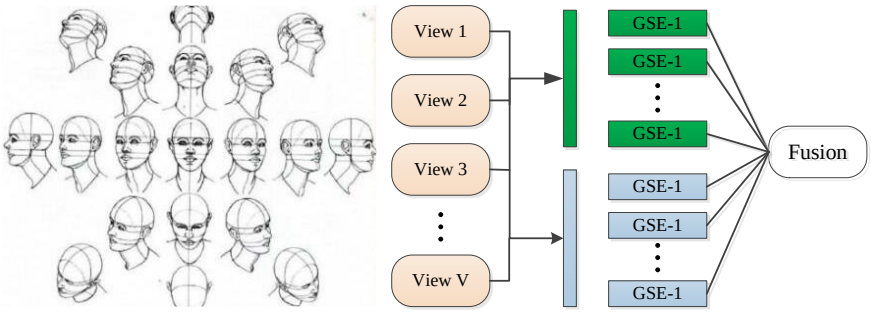


Figure 18: The framework of multi-view clustering model based on spectral embedding fusion (MVSEF) (adapted from [141])

where the variable $\alpha^{(v)}$ denotes the nonnegative normalized parameter for controlling the importance of the *v-th* view.

Although SC is powerful and widely used, researchers attempt to extend it to the deep neural network framework in recent several years [134, 135, 136, 137, 138, 139, 140]. For example, Law *et al*. [134] propose a deep supervised spectral clustering method, in which a metric learning framework is optimized so as to obtain more reasonable clusters. Yang *et al*. [136] propose a joint framework for discriminative embedding and spectral clustering in a dual auto-encoder network. Affeldt *et al*. [138] propose a model which combines spectral clustering and deep auto-encoder (SC-EDAE) strengths in an ensemble framework. Recently, there are also several attempts to explore the effectiveness of spectral clustering on deep multi-view settings, such as spectral embedding fusion [141], multi-view spectral network [142] and multi-view spectral clustering network [143]. Next, we review several representative deep multi-view extensions of spectral clustering method.

3.3.2. Spectral Embedding Fusion

To simultaneously incorporate the local and global structures of multiple views, Yin *et al*. [141] present a MVC model based on spectral embedding fusion, named MVSEF, as in Figure 18. To discover the final fusional embedding, the proposed MVSEF builds an objective function, which is solved by an iteration method via $L_{2,1}$ norm.

Suppose that there are two views $\mathbf{Z}^{(v)}$ and $\mathbf{Z}^{(u)}$, which indicate their global spectral embedding. Kumar *et al*. [129] formulate the disagreement between the two views as follows

$$Dis(\mathbf{Z}^{(v)}, \mathbf{Z}^{(u)}) = \| \frac{K_{\mathbf{Z}^{(v)}}}{\|K_{\mathbf{Z}^{(v)}}\|_F^2} - \frac{K_{\mathbf{Z}^{(u)}}}{\|K_{\mathbf{Z}^{(u)}}\|_F^2} \|_F^2 \tag{26}$$

where $K_{\mathbf{Z}^{(v)}}$ denotes a kernel function for $\mathbf{Z}^{(v)}$. The objective function of the MVC model based on spectral embedding fusion (MVSEF) can be formulated as follows

$$\begin{aligned} &\mathcal{L}_{MVSEF} = \min_{\mathbf{Z}^{(v)}, \mathbf{Z}^{(u)}} Tr(\mathbf{Z}^{(v)T} \mathbf{L}^{(v)} \mathbf{Z}^{(v)}) + Tr(\mathbf{Z}^{(u)T} \mathbf{L}^{(u)} \mathbf{Z}^{(u)}) + \lambda Dis(\mathbf{Z}^{(v)}, \mathbf{Z}^{(u)}), \\ &s.t., \mathbf{Z}^{(v)T} \mathbf{Z}^{(v)} = \mathbf{I}, \mathbf{Z}^{(u)T} \mathbf{Z}^{(u)} = \mathbf{I} \end{aligned} \tag{27}$$

where the variable $\mathbf{L}^{(v)}$ indicates the Laplacian matrix of the *v-th* view. In spectral clustering, global spectral embedding (GSE) can be characterized by the spectral embedding $\mathbf{Z}$. The basic process of MVSEF can be found in Figure 18.

3.3.3. Multi-view Spectral Network

To deeply cluster multi-view data, Huang *et al*. [142] propose multi-view SC network (MvSCN), which can incorporate the consistency across views and the invariance in single view into a joint objective function. Specifically, MvSCN implements deep neural network as $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^c$, where the variable c indicates the cluster number and the d denotes the dimension of the feature representation. Once obtaining the trained parameters θ, k-means can be

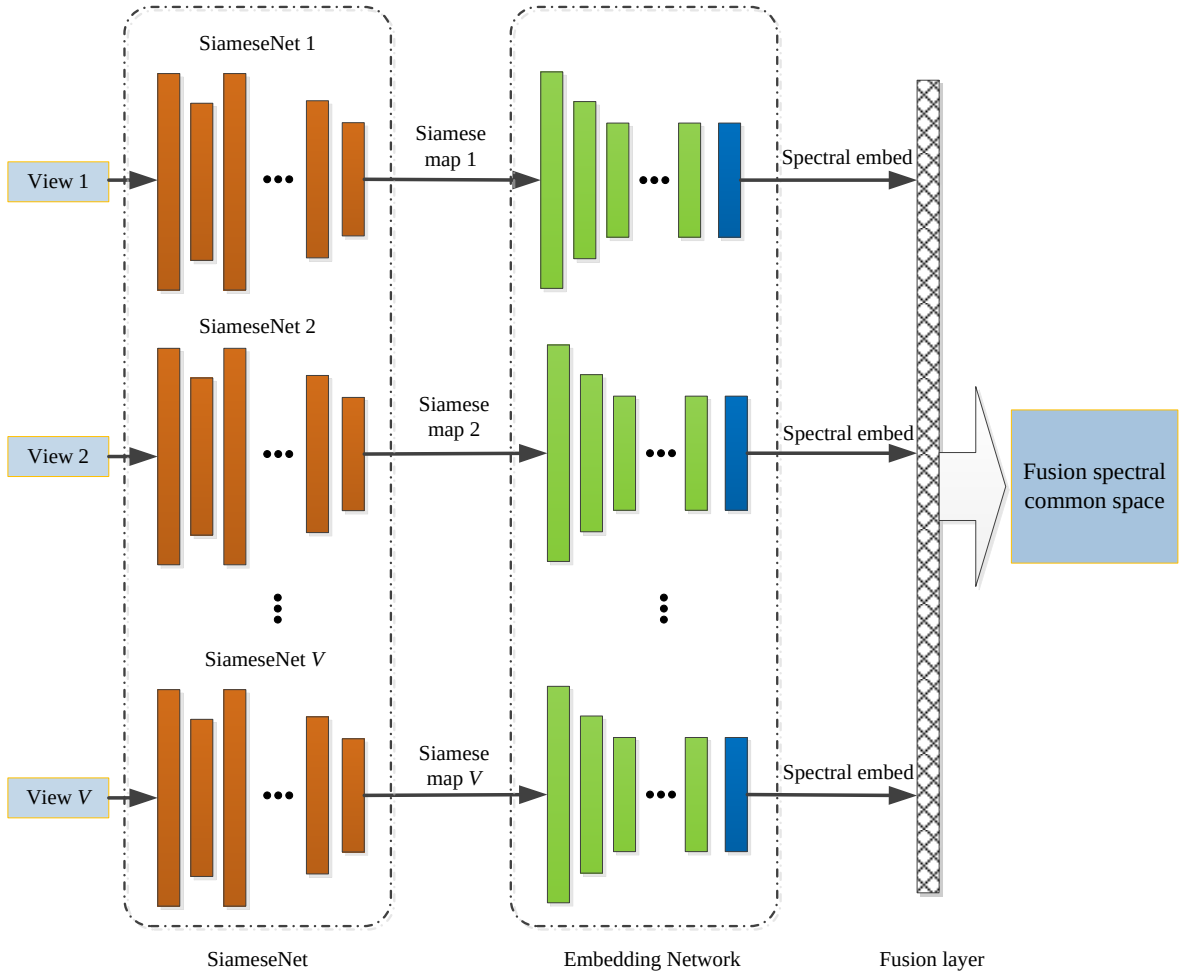


Figure 19: The architecture of the multi-view spectral clustering network (MvSCN) (adapted from [142])

applied to get the cluster assignment. As in [142], the objective function of MvSCN are formulated in two-fold setting as follows

$$\mathcal{L}_{MvSCN} = (1-\lambda)\sum_{v=1}^{V}\mathcal{L}_1^{(v)} + \lambda\mathcal{L}_2^{(v)} \tag{28}$$

where $\lambda \in [0,1]$ denotes a balance parameter to control the trade-off between $\mathcal{L}_1^{(v)}$ and $\mathcal{L}_2^{(v)}$. $\mathcal{L}_1^{(v)}$ is the within-view similarity which enforces the similar points more closer to each other in single view. $\mathcal{L}_2^{(v)}$ is the between-view consistency which minimizes the view discrepancy. The two parts $\mathcal{L}_1^{(v)}$ and $\mathcal{L}_2^{(v)}$ can be formulated as

$$\begin{cases} \mathcal{L}_1^{(v)} = \dfrac{1}{n^2}\displaystyle\sum_{i,j}^{n} W_{ij}^{(v)} \|\mathbf{y}_i^{(v)} - \mathbf{y}_j^{(v)}\|_2^2 \\ \mathcal{L}_2^{(v)} = \dfrac{1}{nm^2}\displaystyle\sum_{v,p}^{m}\sum_{i}^{n} \|\mathbf{y}_i^{(v)} - \mathbf{y}_i^{(p)}\|_2^2 \end{cases} \tag{29}$$

where W is the affinity matrix or graph. $\mathbf{y}_i^{(v)}$ is the output of a neural network of input data $\mathbf{x}_i^{(v)}$. The MvSCN model is trained in a coordinate descent fashion which alternates between the orthogonalization and the gradient steps as

shown in Figure 19. MvSCN model contains the following stages: affinity learning by SiameseNet [144], embedding networks for each view, QR decomposition by orthogonal layer and common space learning by fusion layer.

3.4. Deep Multi-view Information Bottleneck

In this part, the information bottleneck (IB) [106] theory and its multi-view extensions are firstly presented. Then, two typical deep multi-view IB based methods are reviewed.

3.4.1. Information Bottleneck

Given a source variable X and its relevant variable Y, IB method aims to map the source data X into its compressed one T, which is formulated by the mutual information between them, i.e., $I(X;T)$. At the same time, the preserving the relevant information about variable Y maximally is formulated by $I(Y;T)$. Thus, the objective function of IB theory is given as follows

$$\mathcal{L}_{max} = I(Y;T) - \beta^{-1} I(X;T), \tag{30}$$

where β balances the data compression and related information preservation.

In the past decades, lots of multi-view extensions of the traditional IB approaches have been applied into many practical fields. For instance, Gao et al. [145] design a multi-view IB approach through adopting a compatible constraint on different views to attain an consistent clustering result. Yan et al. present multi-feature IB [29, 146], multi-task IB [29, 147, 148], heterogeneous IB [149, 150] for unsupervised categorization, cross-modal clustering, respectively. Yan et al. [30, 151] jointly address the ensemble and multi-view clustering problem by presenting a synergetic IB method. Hu et al. [152, 153] propose a joint view-specific and view-shared information utilization based IB method for human action clustering. However, all the above methods are based on traditional machine learning models, which cannot fully capture the comprehensive feature representations of different views. Therefore, incorporating deep neural networks into the IB principle based multi-view learning framework is still a difficult issue.

In the last several years, the deep models based IB methods [154, 155, 156] are devised and successfully tackle the above challenges. In the followings, we list some typical deep multi-view IB based methods in detail.

3.4.2. Deep Multi-view Variational Information Bottleneck

To remove the noisy and irrelevant information in different views, a deep IB based multi-view learning framework (DMVIB) [157] is propose to learn an accurate shared representation across views. DMVIB aims to maximally preserve the mutual information between the learned shared representation and data labels while compressing the original representation into shared representation as much as possible. Specifically, DMVIB mainly has two parts, (1) learning a hidden representation for each view; (2) combining these hidden representations to learn a shared one across views.

Given m views $\{X_i\}_{i=1}^m$ and their corresponding labels Y, DMVIB aims to discover a shared feature space Z among views. First, to remove the noisy and irrelevant information in individual views, a hidden representation H_i is learned for the i-th view. Then, a neural work is adopted to fuse these hidden representations as follows

$$Z = f_\theta(H_1, H_2, \ldots, H_m), \tag{31}$$

where f indicates the neural network and θ is the parameters.

Thus, the above two parts are formulated as

$$\max_{Z,H_1,H_2,\ldots H_m} I(Y,Z) - \sum_{i}^{m} \alpha_i I\left(X_i, H_i\right), \tag{32}$$

where α_i denotes the regularization parameter trading off the two parts.

The two terms balance the model accuracy and complexity and make it a more generalizable model. To optimize the above objective function, variational inference approach can be adopted and more details are seen in the work [157].

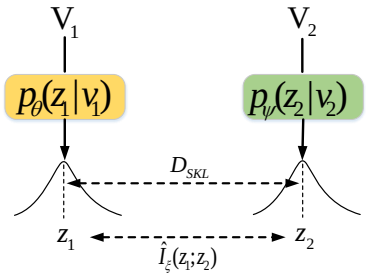


Figure 20: Unsupervised deep multi-view information bottleneck (MIB) method.

3.4.3. Robust Representation Learning via Multi-View Information bottleneck

Different from the above supervised deep multi-view IB method [157], Federici et al. propose an unsupervised deep multi-view IB method (MIB) [158] for learning robust representations among views by discovering the superfluous information which is not shared by different views. This work is based on a basic assumption in multi-view community– each view has the same task-relevant information. Thus, an improved robust and generalized representation can be learned by abandoning the information not shared by views, i.e., removing the view-specific nuisances. This can be realized by designing a multi-view InfoMax objective to maximizing the mutual information among views.

Given v_1 and v_2 denoting two images of one specific object from two views. Let z_1 and z_2 denote the representations of v_1 and v_2 respectively. Let z denote the latent representation that only contain the necessary information for predicting data labels. To realize sufficiency for predicting the data label, the representation z_1 of v_1 should be ensured to be sufficient for v_2. In the meanwhile, decreasing the $I(z_1; v_1|v_2)$ is also needed to ensure robustness of the learned representations. Thus, the objective function of MIB can be written as follows

$$\mathcal{L}_{MIB}(\theta, \psi; \beta) = -I_{\theta\psi}(z_1; z_2) + \beta D_{SKL}\left(p_\theta(z_1 \mid v_1) \| p_\psi(z_2 \mid v_2)\right), \tag{33}$$

where θ indicates the parameters of encoder $p_\theta(z_1|v_1)$, ψ indicates the parameters of encoder $p_\psi(z_2|v_2)$, β is the trade-off parameter balancing between sufficiency and robustness of the learned representation, D_{SKL} indicates the symmetric KL divergence by imposing average on the values of $D_{\mathrm{KL}}\left(p_\theta(z_1 \mid v_1) \| p_\psi(z_2 \mid v_2)\right)$ and $D_{\mathrm{KL}}\left(p_\psi(z_2 \mid v_2) \| p_\theta(z_1 \mid v_1)\right)$. For better understanding, the MIB method is further illustrated in Figure 20.

The proposed MIB method can well address some practical applications with pairwise observations available or synthetic, such as image retrieval. However, MIB can only deal with two-view representation learning problems, and an extended version is imperative for application into more complicated scenarios. In big data era, there emerges large amounts of data from various fields every day. Almost all of them are unlabelled and it is impractical to label them for training of supervised models. Although the above MIB method has shown its effectiveness in unsupervised learning, it still has many limitations and many possible future interests can be touched. Therefore, more unsupervised deep multi-view IB methods are worth exploring in the future. Additionally, recent deep multi-view IB methods only utilize mutual information to measure the relationships among views, and the correlated information may be not well propagated across views, which may degrade the final classification performance. Hence, more effective and efficient deep IB-based propagating strategies require to be designed in the future. Finally, lots of challenges faced by many complex practical applications, such as deep multi-view action recognition and image classification, are still unresolved, where deep multi-view IB method may probably provide a possible solution.

4. Applications

In the past decades, deep MVL methods have obtained promising results in computer vision and pattern recognition communities due to their remarkable abilities in feature representations, such as object retrieval, social video/image analysis, bioinformatics and health informatics, natural language processing, and recommendation systems. Among those, the former three fields are quite prosperous in the very recent years and thus we conduct a detailed survey on them in the following.

4.1. Cross-modal Retrieval

Cross-modal retrieval (CMR) is a fundamental research topic which intends to search the data instances from other modalities with similar semantic, such as using a text to retrieve the related images [90]. Recently, the cross-modal retrieval based deep learning mode has achieved great progress. The existing deep cross-modal retrieval methods can be usually classified into two types: Hashing and subspace methods.

The hashing based deep CMR methods aim to find a hash function enabling transformation between different modalities. Deep cross-modal hashing approaches involve two settings: unsupervised and supervised approaches. Specifically, the unsupervised ones aim to learn hashing functions by exploring the modality information of the unlabeled source data, such as deep joint-semantics reconstructing hashing (DJSRH) [90], unsupervised deep cross-modal hashing (UDCMH) [159], self-supervised adversarial hashing (SSAH) [160] and unsupervised coupled hashing (UCH) [161]. The supervised ones aim to exploit available supervised information (like labels) to improve retrieval performance, such as graph convolutional hashing (GCH) [162], triplet based deep hashing (TDH) [163], cycle-consistent deep generative hashing (CYC-DGH) [164] and equally-guided discriminative hashing (EGDH) [165].

The common subspace based CMR methods try to discover a common feature space, where the distance between different modalities can be directly computed. For example, Zhen *et al*. [166] propose a deep supervised CMR (DSCMR) method, which minimally preserves the discrimination loss in both the subspace and the label space to supervise the common feature learning of multiple modalities. Huang *et al*. [167] propose a framework leveraging multi-modal neural machine translation (MMT) by performing forward and backward translations of multiple languages.

4.2. Cross-modal Video/Image Analysis

Deep MVL methods have been successfully applied to the analysis of cross-modal images or videos. Next, we will demonstrate some representative ones.

4.2.1. 3D Reconstruction

Due to the promising performance and widely application of DNNs, 3D reconstruction approaches based learning algorithm have achieved popularity [168, 169, 60]. Dou et al. [168] present a deep RNN model for multi-view 3D face reconstruction to overcome the main issue of huge variations in facial images. Instead of using RNN model, Yang et al. [60] design a new feed-forward neural network and a dedicated training method to gather deep features for the 3D object reconstruction from multiple views, so that the issue of inconsistent estimation of 2D shapes under permutations of input images can be avoided.

4.2.2. Facial Detection and Recognition

Facial detection from multi-view source data is a valuable research topic and challenging task since the there exists large viewpoint changes in poses and illumination [170, 171, 173, 174, 176, 177]. To deal with the problem of face detection in MVL scenario, Farfade et al. [170] propose a deep CNN based dense face detector without requiring annotation of facial landmarks and training multiple models for capturing faces in different orientations. Bai et al. [171] formulate the 3D face detection problem from the point of non-rigid multi-view stereo, and thus presented to optimize the 3D face shape by keeping the multi-view appearances consistent. Li et al. [172] aim to solve the large-scale face recognition problem by proposing a robust two-stage method containing data cleaning and multi-view deep representation learning. Xia et al. [175] intend to estimate the eye center using a CNN network. Zhao et al. [176] formulate a novel DNN architecture to improve the performance of multi-view face recognition by first encoding the face regions, compressing the high-dimensional learned features and designing a joint Bayesian framework for classification.

4.2.3. Human Action Recognition

Human actions in videos or images usually consist of highly articulated motions, human-object interactions and complicated temporal structures [178, 179, 180, 181, 182]. Song et al. [178] present a new action recognition framework under multi-modal scenarios based on deep CNN and RNN architectures and can better learn effective feature representations for action classification. Classifying imbalanced multi-modal sensor data in the environment of smart home for activity recognition is quite challenging. To this end, Alani et al. [179] first examine the effectiveness of

using multi-modal data and then compare deep learning methods with other methods in addressing the imbalanced multi-modal data. Trumble et al. [180] propose a deep CNN based human performance capture system for the challenging marker-less pose estimation from multi-view videos. Huang et al. [181] present a two-stage 3D neural network for estimating 3D human pose by combining both body-worn inertial measurement unit (IMU) data and cross-view images, where the first stage is used for vision estimation and the second stage for fusing the early IMU data and vision data without the need of a skeleton model.

4.2.4. Person Re-identification

Object or person re-identification has received lots of attention in the image processing community, and we here show two typical works for each of them [183, 184, 185]. For instance, Zhou et al. [184] propose to integrating CNN and RNN architecture for the re-identification of vehicle in arbitrary view point, so that the transformations among vehicles across different views are well learned for improving performance. Instead of using semi-supervised learning for person re-identification, Xin et al. [185] focus on incorporating self-spaced learning into a multi-view clustering paradigm, where the reliable data points are used for fine-tuning the CNN model by introducing a novel regularizer for minimizing both the identification loss and ranking loss.

4.3. Bioinformatics and Health Informatics

Due to the success of deep models in the task of feature representation, lots of popular DNN based networks (e.g., CNN) have been applied to the challenging but beneficial medical analysis [186, 187, 188, 189, 190], such as bioinformatics and health informatics. More specifically, it has been widely used in the three typical areas including brain issues [8, 9], breast diagnosis [191, 192, 193] and seizure recognition [194, 195]. For instance, Wei et al. [8] propose a deep neural network architecture for brain image segmentation in the multi-modal/size/view scenarios, such that the coronal or sagittal MR slices can be clearly segmented. For magnification invariant diagnosis in breast cancer, Jonnalagedda et al. [192] propose a multi-view path DNN and a data augmentation method to deal with the challenging issues of diversity and small size of datasets and it further integrates local and global features for more effective diagnosis. By combining two popular models of DNN and SVM, Gong et al. [193] present a multi-view DNN architecture based SVM classification method for breast cancer diagnosis under the modality of B-mode ultrasound and ultrasound electrography. Yuan et al. [194] try to detect epileptic seizure by using multi-channel scalp electroencephalogram signals with a multi-view deep channel-aware attention network. To improve the model, Yuan et al. [195] further design a enhanced end-to-end model for joint unsupervised electroencephalogram reconstruction and supervised seizure detection.

There are also some other emerging medical applications of deep multi-view learning methods, such as cardiac magnetic resonance detection [196], microscopic neuroblastoma pathology image diagnosis [197], and automated diagnosis of bone metastasis [198].

5. Datasets

In the following, we list some popular multi-view text, image and video datasets and show the details in Table 2. Note that, for image datasets, each kind of feature, such as shape, texture or color, is always treated as one view, so we do not show them in the table. Here are some popular image datasets used for evaluating multi-view learning methods: Caltech 101/256 [199], NUS-WIDE [200], COIL20/100 [1], 17flowers [201], Soccer [2], Scene-15 [202], and ORL [3]datasets.

Table 2: Details about the representative multi-view datasets.

[1] https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
[2] http://lear.inrialpes.fr/people/vandeweijer/data
[3] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

Dataset	Type	Views	Categories	Size	Year	Highlights
20NGs [4]	Text	3	5	500	-	Selected from the popular 20Newsgroups dataset; Documents are randomly divided into 3 sections with the same number of documents, where each part is regarded as one view.
continued on next page						

continued from previous page

Dataset	Type	Views	Categories	Size	Year	Highlights
Reuters [203]	Text	5	6	1,200	2009	Typical multi-language text dataset; Documents written in five types of languages (German, English, French, Italian and Spanish), where each language is treated as one view.
BBC [5]	Text	4	5	2,225	2006	Popular multi-view text dataset; Story documents of five topical fields from BBC website in 2004-2005.
BBC Sport [6]	Text	2	5	737	2006	It is a typical multi-view text dataset, in which the text documents are recorded from the BBC news website happened in 2004-2005.
3Sources [7]	Text	3	6	169	2009	Representative multi-view text dataset; News articles concerning stories from different areas from the BBC Reuters and Guardian website.
Cora [8]	Text	2	7	2,708	-	Typical multi-view publication text dataset; Scientific publications described in "Content" and "Citation" view.
WVU [204]	Video	4	10	650	2011	Ten kinds of human actions captured by 8 embedded camera networks; Complete overlapping coverage is provided for 8 different perspectives/views;
IXMAS[205]	Video	5	10	1,320	2007	It is a video dataset containing 11 human actions, the actions of each category is conducted 3 times by 10 actors. In the recording procedure, all actions are recorded by 5 cameras from different viewpoints, which consists of 4 side-viewing and 1 top-viewing camera.
M^2I[206]	Video	2	22	1,760	2017	Typical challenging multi-view action dataset; Multi-modal and multi-view interactive human actions collected in dark, bright, and cluttered environments; Modalities, including RGB, depth, and skeleton, are simultaneously captured from front and side views by two static Kinect depth sensors;
MV-10K[207]	Video	2	10	10,000	2018	It is a popular audio-visual dataset, in which the lasting time of each video varies from 213 to 219 seconds.
VEGAS [208]	Video	2	10	28,103	2018	Selected from Google Audioset, containing human or animal sounds. Each video ranges from 2 to 10 seconds.
MSR 3D[209]	Video	2	16	320	2012	Popular multi-modal dataset (RGB, depth, and skeleton); Each action is performed in two different poses, most of which involve human-object interaction.

continued on next page

continued from previous page

Dataset	Type	Views	Categories	Size	Year	Highlights
N-UCLA[210]	Video	3	10	385	2014	Captured simultaneously from different views by three Kinect cameras; Each action sample contains RGBD and human skeleton data executed by 10 different subjects.
MVTJU[211]	Video	2	22	3,520	2015	Each action was performed 4 times by 20 subjects under two lighting conditions (i.e., light and dark); Each action was recorded simultaneously using two Kinect depth sensors (frontal and lateral views), including RGB, depth, and skeletal data.
Middlebury [9]	3D	3	5	100	2011	This dataset provides multiple datasets capturing objects from various points.
ModelNet40 [10]	3D	12	40	12,311	2015	It is a dataset containing different 3D models, in which each 3D model can be partitioned into a 30×30×30 shape.
KITTI [11]	3D	4	389	14,999	2012	This dataset contains the object detection part published for autonomous driving. It contains a set of images with their bounding box labels.
NUS-WIDE [12]	Cross-modal	2	81	270,000	2009	It is a cross-model dataset containing visual images and its tags from 81 categories.
BDGP [13]	Cross-modal	2	5	2,500	2012	It is a cross-modal dataset consisting of two different modalities, i.e., visual image and textual semantic data.
COCO [14]	Cross-modal	2	80	123,287	2014	It is a dataset constructed for segmentation, object detection and captioning dataset, which includes two different views, i.e., visual object and its context.
FLICKR [15]	Cross-modal	2		25,000	2008	It is a collection for the MIR community comprising 25000 images from the Flickr website, which contains both image content and image tags.

6. Performance Comparisons

In the last several years, a large variety of deep multi-view learning approaches has appeared and the collections of standard benchmarks such as Reuters, MNIST and NUS-Wide have made it easier to compare these deep multi-view

[4] http://lig-membres.imag.fr/grimal/data.html
[5] http://mlg.ucd.ie/datasets/bbc.html
[6] http://mlg.ucd.ie/datasets/
[7] http://mlg.ucd.ie/datasets/3sources.html
[8] https://relational.fit.cvut.cz/dataset/CORA
[9] https://vision.in.tum.de/data/datasets/3dreconstruction
[10] https://shapenet.cs.stanford.edu/media/modelnet40_normal_resampled.zip
[11] www.cvlibs.net/datasets/kitti
[12] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm
[13] http://ranger.uta.edu/%7ehent/Drosophila/
[14] http://mscoco.org/
[15] http://press.liacs.nl/mirflickr/mirdownload.html

learning approaches in terms of accuracy. Although it may be impractical to conduct comparison on all proposed approaches, it is necessary to compare typical deep MVL approaches in a unified manner.

In this part, we compare the performance of deep MVL approaches on three common tasks: cross-modal retrieval, multi-view 3D object recognition and multi-view clustering (MVC). The MVL is concerned with the problem of machine learning from data represented by multiple distinct features or data sources, which is a learning method in essence. Thus, the common performance indicators in the domain of machine learning can be utilized to evaluate the performance of MVL methods. For the deep cross-modal retrieval methods, we compare the mean average precision (mAP) [162] of the following approaches: deep joint-semantics reconstructing hashing (DJSRH) [90], self-supervised adversarial hashing (SSAH) [160], unsupervised deep cross-modal hashing (UDCMH) [159], unsupervised coupled hashing (UCH) [161], graph convolutional hashing (GCH) [162], triplet based deep hashing (TDH) [163], cycle-consistent deep generative hashing (CYC-DGH) [164], equally-guided discriminative hashing (EGDH) [165], deep supervised cross-modal retrieval (DSCMR) [166] and multi-modal neural machine translation (MMT) [167]. For the deep multi-view object recognition methods, we compare the classification accuracy (Accuracy) [49] of the following approaches: multi-view CNN (MVCNN) [49], Kd-tree based network (KD-Net) [51], multi-view harmonized bilinear network (MHBN) [52], group-view CNN (GVCNN) [50] and dynamic routing CNN (DRCNN) [53] For deep MVC methods, we compare the clustering accuracy (ACC) and normalized mutual information (NMI) [142] of the follow representative approaches: multi-view spectral clustering network (MvSCN) [142], auto-encoder in auto-encoder network (AE^2-Nets) [59], spectral clustering and deep auto-encoder (SC-EDAE) [138], auto-weighted multi-view learning (AWMVL) [116], cross-modal auto-encoders (CMAE) [19], joint NMF (JNMF) [118], bidirectional GAN (BiGAN) [36], MultiSpectralNet (MSN) [143], adversarial correlated auto-encoder (ACAE) [69], deep multi-view concept learning (DMCL) [122], deep CCA auto-encoder (DCCAE) [22] and deep multi-view IB (MIB) [158].

Table 3: The performance comparison of representative deep cross-modal retrieval methods

NUS-WIDE		COCO		Flickr	
Methods	Map score	Methods	Map score	Methods	Map score
DSCMR [166]	0.615	UCH [161]	0.547	UCH [161]	0.697
TDH [163]	0.676	SSAH [160]	0.578	UDCMH [159]	0.733
SSAH [160]	0.683	EGDH [165]	0.813	TDH [163]	0.755
UDCMH [159]	0.761	GCH [162]	0.830	CYC-DGH [164]	0.799
GCH [162]	0.766	–	–	MMT [167]	0.801
DJSRH [90]	0.817		–	DJSRH [90]	0.876
–	–	CYC-DGH [164]	0.895	GCH [162]	0.907

Table 4: The performance comparison of representative deep multi-view object recognition methods

ModelNet40		ModelNet10	
Methods	Accuracy	Methods	Accuracy
MVCNN [49]	0.899	MVCNN [49]	0.927
KD-Net [51]	0.907	KD-Net [51]	0.940
MHBN [52]	0.934	MHBN [52]	0.950
GVCNN [50]	0.926	–	–
DRCNN [53]	0.945	DRCNN [53]	0.960

7. Open Problems

7.1. Explainable Deep Multi-view Model

Explainable artificial intelligence has become a newly-emerging area in recent years and has achieved lots of attention in the field of deep MVL. Although existing deep MVL models have shown superior advantages in various applications, they fail to provide an explanation for the decision of different models. Thus, the deep models without

Table 5: The performance comparison of representative deep multi-view clustering methods

MNIST			Reuters			Handwritten		
Methods	ACC	NMI	Methods	ACC	NMI	Methods	NMI	ACC
MvSCN [142]	0.991	0.977	CRSC [129]	–	0.361	AE^2-Nets [59]	0.815	0.713
SC-EDAE [138]	0.915	–	MvSCN [142]	0.488	0.267	AWMVL [116]	0.973	0.939
CMAE [19]	0.911	–	JNMF [118]	0.535	0.409	CRSC [129]	–	0.768
BiGAN [36]	0.973	–	MSN [143]	0.631	0.598	JNMF [118]	0.881	0.804
ACAE [69]	0.927	–	DCML [122]	0.732	–	–	–	–
DCCAE [22]	0.975	0.934	MCL [123]	0.761	–	–	–	–
MIB[158]	0.978	–	–	–	–	–	–	–

explanation are not easily applied to critical domains, such as military system or medical treatment. In the future, we believe that explainable deep models will become a hot topic and will be extended into more areas. Recent years, many researches attempt to open the black box of deep neural networks and propose a various of theories to understand it. Among them, information bottleneck (IB) theory claims that there are two distinct phases consisting of fitting phase and compression phase in the course of training [91]. This statement attracts many attentions since its success in explaining the inner behavior of feedforward neural networks.

7.2. Incomplete Views

With the development of information technology, there always emerges data from different views or modalities. However, it happens often that there is usually data missing in many practical applications. For example, in medical imaging analysis, the medical records of patients, e.g., computed tomography or magnetic resonance images, are not always complete and can not fully reflect the conditions of patients, which may further influence the quality of healthcare. Therefore, the completeness of such views and the applications in some specific downstream tasks, e.g., multi-view feature learning, clustering or classification, is significant and it is worth exploring this issue in future. Recently, multi-view transformation with generative adversarial network (GAN) [64] has obtained considerable attention, which is associated with transforming available source views of a given object into unknown or incomplete target views. We think it is interesting topic in the task of view completion.

7.3. Heterogeneous Gap of Different Views

Data points captured from multiple views always hold heterogeneous features but meanwhile illustrating complex view relationships. Thus, it is quite challenging to explore the shared information among different views and the view-specific information simultaneously. Existing methods usually solve this problem in a traditional manner by using k-means or spectral based multi-view learning methods, while few of them touch the deep learning models. Therefore, designing a reasonable deep model by jointly breaking the heterogeneous gap across views and exploring the relationships will be a promising research interest. Actually, there are several practical strategies from heterogeneous data, such as transfer learning and knowledge graph based representation learning. Specifically, transfer learning aims to share information from the source domain to the target domain, where both domains hold heterogeneous feature spaces, such as the case of cross-media intelligence [35]. Thus, when reconstructing feature representation from heterogeneous data, transfer learning helps to bridge the gaps of data distribution, feature space etc., so that strong and robust feature vectors can be well constructed.

7.4. View Relation Exploration

In multi-view clustering, since the same data samples are described by multiple views from different perspectives, different views must be closely related. However, the relationships among views are quite complex and hard to be explored. Most existing deep multi-view learning methods usually learn the shared embedding features to discover the view relations by simple concatenation or fusion, and then apply cross-entropy loss function or traditional machine learning methods for supervised or unsupervised learning. These methods, however, still fail to fully explore the view relations. Hence, designing an effective view relation exploration strategy for deep multi-view learning will be quite significant yet challenging in the future. We argue that the information propagation mechanism may be

borrowed to propagate useful relations among views for relation exploration, such as affinity propagation and label propagation [80].

7.5. Representation Learning on Multi-view Graph Data

Recently, more and more multi-view data exhibit complex graph structures, where one of the most important problems to be handled is the graph feature representation. A good feature representation of the multi-view graph data is important for lots of downstream applications, such as classification/clustering, object segmentation, object detection, or 3D reconstruction. However, only a few works are designed to capture the graph features of multi-view source data from various practical fields, such as social network or medical diagnosis data. In the next few years, the multi-view graph representation will be a hot topic and shows its advantages in different application areas. Recently, Graph neural networks (GNN) [73] reconciles the expressive power of graphs in modeling interactions with deep models in terms of learning representation and has gained increasing attention due to its capability of modeling graph structured data. Using GNN to model the complex multi-view graph data is interesting topic that worth of investigating.

8. Conclusions

This paper presents a comprehensive review on deep MVL from the following two aspects: MVL approaches in deep learning scope and deep multi-view extensions of traditional learning methods. Specifically, we first review the representative MVL methods in the scope of deep learning, such as multi-view auto-encoder, conventional neural networks and deep brief networks. Then, we explore the advancements of the MVL mechanism when traditional learning methods meet deep learning models, such as deep multi-view canonical correlation analysis, matrix factorization and information bottleneck. Moreover, we also summarize the main applications and widely-used datasets in the domain of deep MVL. Finally, to promote the research of deep MVL, we present several open challenges that are worth further investigation in future.

References

[1] H. Guo, J. Wang, M. Xu, Z. Zha, H. Lu, Learning multi-view deep features for small object retrieval in surveillance scenarios, in: Proceedings of the Annual ACM Conference on Multimedia Conference (ACM'MM), 2015, pp. 859–862.

[2] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1933–1941.

[3] K. Deepak, G. Srivathsan, S. Roshan, S. Chandrakala, Deep multi-view representation learning for video anomaly detection using spatiotemporal autoencoders, Circuits, Systems, and Signal Processing (2020) 1–17.

[4] N. Srivastava, R. Salakhutdinov, Multimodal learning with deep boltzmann machines, Journal of Machince Learning Research (JMLR) 15 (1) (2014) 2949–2980.

[5] J. Mao, W. Xu, Y. Yang, J. Wang, A. L. Yuille, Deep captioning with multimodal recurrent neural networks (m-rnn), in: Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[6] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, IEEE Transactions on Pattern Analysis and Machince Intelligence (TPAMI) 39 (4) (2017) 664–676.

[7] W. Fan, Y. Ma, H. Xu, X. Liu, J. Wang, Q. Li, J. Tang, Deep adversarial canonical correlation analysis, in: Proceedings of the SIAM International Conference on Data Mining (SDM), 2020, pp. 352–360.

[8] J. Wei, Y. Xia, Y. Zhang, M^3net: A multi-model, multi-size, and multi-view deep neural network for brain magnetic resonance image segmentation, Pattern Recognition (PR) 91 (2019) 366–378.

[9] J. Xu, H. Zheng, J. Wang, D. Li, X. Fang, Recognition of EEG signal motor imagery intention based on deep multi-view feature learning, Sensors 20 (12) (2020) 3496.

[10] S. Sun, A survey of multi-view machine learning, Neural Computation Applications 23 (7-8) (2013) 2031–2038.

[11] J. R. KETTENRING, Canonical analysis of several sets of variables, Biometrika 58 (3) (1971) 433–451.

[12] Y. Zhang, J. Zhang, Z. Pan, D. Zhang, Multi-view dimensionality reduction via canonical random correlation analysis, Frontiers Computer Science 10 (5) (2016) 856–869.

[13] L. Sun, B. Ceran, J. Ye, A scalable two-stage approach for a class of dimensionality reduction techniques, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2010, pp. 313–322.

[14] H. Avron, C. Boutsidis, S. Toledo, A. Zouzias, Efficient dimensionality reduction for canonical correlation analysis, in: Proceedings of the 30th International Conference on Machine Learning (ICML), Vol. 28, 2013, pp. 347–355.

[15] X. Zhang, X. Yang, W. Zhang, G. Li, H. Yu, Crowd emotion evaluation based on fuzzy inference of arousal and valence, Neurocomputing, 2021

[16] A. Kumar, H. D. III, A co-training approach for multi-view spectral clustering, in: Proceedings of the International Conference on Machine Learning (ICML), 2011, pp. 393–400.
[17] Z. Xue, J. Du, D. Du, S. Lyu, Deep low-rank subspace ensemble for multi-view clustering, Information Sciences 482 (2019) 210–227.
[18] F. R. Bach, M. I. Jordan, Kernel independent component analysis, Journal of Machine Learning Research (JMLR) 3 (2002) 1–48.
[19] G. Bhatt, P. Jha, B. Raman, Representation learning using step-based deep multi-modal autoencoders, Pattern Recognition (PR) 95 (2019) 12–23.
[20] G. Andrew, R. Arora, J. A. Bilmes, K. Livescu, Deep canonical correlation analysis, in: Proceedings of the International Conference on Machine Learning (ICML), 2013, pp. 1247–1255.
[21] Y. Ming, X. Meng, C. Fan, H. Yu, Deep learning for monocular depth estimation: A review, Neurocomputing Volume 438, 28 May 2021, Pages 14-33
[22] W. Wang, R. Arora, K. Livescu, J. A. Bilmes, On deep multi-view representation learning, in: Proceedings of the International Conference on Machine Learning (ICML), 2015, pp. 1083–1092.
[23] P. L. Lai, C. Fyfe, Kernel and nonlinear canonical correlation analysis, International Journal of Neural Systems 10 (5) (2000) 365–377.
[24] S. Akaho, A kernel method for canonical correlation analysis, in: Proceedings of the International Meeting of the Psychometric Society (IMPS), 2001.
[25] R. Socher, F. Li, Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 966–973.
[26] G. Chao, S. Sun, J. Bi, A survey on multi-view clustering (2017).
URL `http://arxiv.org/abs/1712.06246`
[27] C. Xu, D. Tao, C. Xu, A survey on multi-view learning (2013).
URL `http://arxiv.org/abs/1304.5634`
[28] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: Recent progress and new challenges, Information Fusion 38 (2017) 43–54.
[29] X. Yan, Y. Ye, Z. Lou, Unsupervised video categorization based on multivariate information bottleneck method, Knowledge-Based Systems (KBS) 84 (2015) 34–45.
[30] X. Yan, Y. Ye, X. Qiu, H. Yu, Synergetic information bottleneck for joint multi-view and ensemble clustering, Information Fusion 56 (2020) 15–27.
[31] Y. LeCun, Y. Bengio, G. E. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[32] J. Wei, Y. Xia, Y. Zhang, M^3net: A multi-model, multi-size, and multi-view deep neural network for brain magnetic resonance image segmentation, Pattern Recognition (PR) 91 (2019) 366–378.
[33] L. Tran, X. Yin, X. Liu, Disentangled representation learning GAN for pose-invariant face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1283–1292.
[34] X. He, Q. Liu, Y. Yang, MV-GNN: multi-view graph neural network for compression artifacts reduction, IEEE Transactions on Image Processing (TIP) 29 (2020) 6829–6840.
[35] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, in: Proceedings of the ACM International Conference on Multimedia ACM'MM, 2014, pp. 7–16.
[36] J. Donahue, P. Krähenbühl, T. Darrell, Adversarial feature learning, in: Proceedings of the International Conference on Learning Representations (ICLR), 2017.
[37] D. Ramachandram, G. W. Taylor, Deep multimodal learning: A survey on recent advances and trends, IEEE Signal Processing Magazine (ISPM) 34 (6) (2017) 96–108.
[38] T. Baltrusaitis, C. Ahuja, L. P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 41 (2) (2019) 423–443.
[39] Y. Li, M. Yang, Z. Zhang, A survey of multi-view representation learning, IEEE Transactions on Knowledge and Data Engineering (TKDE) 31 (10) (2019) 1863–1883.
[40] W. Guo, J. Wang, S. Wang, Deep multimodal representation learning: A survey, IEEE Access 7 (2019) 63373–63394.
[41] N. Zeng, Z. Wang, W. Liu, H. Zhang, K. Hone, X. Liu, A dynamic neighborhood-based switching particle swarm optimization algorithm, IEEE Transactions on Cybernetics (TCYB) 1 (1) (2020) 1–12.
[42] W. Liu, Z. Wang, Y. Yuan, N. Zeng, K. Hone, X. Liu, A novel sigmoid-function-based adaptive weighted particle swarm optimizer, IEEE Transactions on Cybernetics (TCYB) 1 (1) (2019) 1–10.
[43] I. U. Rahman, Z. Wang, W. Liu, B. Ye, M. Zakarya, X. Liu, An n-state markovian jumping particle swarm optimization algorithm, IEEE Transactions on Systems, Man, and Cybernetics: Systems (TSMC) 1 (1) (2020) 1–13.
[44] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, T. Pham, Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation 20 (10) (2020) 2325–2332.
[45] S. Liu, Y. Xia, Z. Shi, H. Yu, Z. Li, J. Lin, Deep Learning in Sheet Metal Bending With a Novel Theory-Guided Deep Neural Network, IEEE/CAA Journal of Automatica Sinica vol. 8, no. 3, pp. 565-581, March 2021.
[46] Z. Yang, L. Tang, K. Zhang, P. Wong, Multi-view CNN feature aggregation with ELM auto-encoder for 3d shape recognition, Cognitive Computation 10 (6) (2018) 908–921.
[47] K. Liu, G. Kang, 3d multi-view convolutional neural networks for lung nodule classification, Plos One 12 (1) (2017) 12–22.
[48] R. Mane, N. Robinson, A. P. Vinod, S. Lee, C. Guan, A multi-view CNN with novel variance layer for motor imagery brain computer interface, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2020, pp. 2950–2953.
[49] H. Su, S. Maji, E. Kalogerakis, E. G. Learned-Miller, Multi-view convolutional neural networks for 3d shape recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 945–953.
[50] Y. Feng, Z. Zhang, X. Zhao, R. Ji, Y. Gao, GVCNN: group-view convolutional neural networks for 3d shape recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 264–272.
[51] R. Klokov, V. S. Lempitsky, Escape from cells: Deep kd-networks for the recognition of 3d point cloud models, in: Proceedings of the IEEE

International Conference on Computer Vision (ICCV), 2017, pp. 863–872.
[52] T. Yu, J. Meng, J. Yuan, Multi-view harmonized bilinear network for 3d object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 186–194.
[53] K. Sun, J. Zhang, J. Liu, R. Yu, Z. Song, DRCNN: dynamic routing convolutional neural network for multi-view 3d object recognition, IEEE Transactions on Image Processing (TIP) 30 (2021) 868–877.
[54] Q. Dou, H. Chen, L. Yu, J. Qin, P. Heng, Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection, IEEE Transactions on Biomedical Engineering (ITBE) 64 (7) (2017) 1558–1567.
[55] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, IEEE Transactions on Image Processing (TIP) 24 (12) (2015) 5659–5670.
[56] Z. Zhang, Q. Zhu, G. Xie, Y. Chen, Z. Li, S. Wang, Discriminative margin-sensitive autoencoder for collective multi-view disease analysis, Neural Networks 123 (2020) 94–107.
[57] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, Multimodal deep learning, in: Proceedings of the International Conference on Machine Learning (ICML), 2011, pp. 689–696.
[58] P. Vincent, H. Larochelle, Y. Bengio, P. Manzagol, Extracting and composing robust features with denoising autoencoders, in: Proceedings of the International Conference on Machine Learning (ICML), Vol. 307, 2008, pp. 1096–1103.
[59] C. Zhang, Y. Liu, H. Fu, Ae2-nets: Autoencoder in autoencoder networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2577–2585.
[60] B. Yang, S. Wang, A. Markham, N. Trigoni, Robust attentional aggregation of deep feature sets for multi-view 3d reconstruction, Internationa Journal of Computer Vision (IJCV) 128 (1) (2020) 53–73.
[61] W. Yan, Detecting gas turbine combustor anomalies using semi-supervised anomaly detection with deep representation learning, Cognitive Computation 12 (2) (2020) 398–411.
[62] F. Liu, Z. Wang, Automatic "ground truth" annotation and industrial workpiece dataset generation for deep learning, International Journal of Automation and Computing (IJAC) 17 (4) (2020) 539–550.
[63] W. Zheng, L. Yan, C. Gou, F. Wang, KM4: Visual reasoning via Knowledge Embedding Memory Model with Mutual Modulation, Information Fusion, Volume 67, March 2021, Pages 14-28.
[64] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS), 2014, pp. 2672–2680.
[65] P. Isola, J. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5967–5976.
[66] B. Dolhansky, C. Canton-Ferrer, Eye in-painting with exemplar generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7902–7911.
[67] Y. Tian, X. Peng, L. Zhao, S. Zhang, D. N. Metaxas, CR-GAN: learning complete representations for multi-view generation, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 942–948.
[68] R. Huang, S. Zhang, T. Li, R. He, Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2458–2467.
[69] X. Wang, D. Peng, P. Hu, Y. Sang, Adversarial correlated autoencoder for unsupervised multi-view representation learning, Knowledge-Based Systems (KBS) 168 (2019) 109–120.
[70] Q. Xuan, Z. Chen, Y. Liu, H. Huang, G. Bao, D. Zhang, Multiview generative adversarial network and its application in pearl classification, IEEE Transactions on Industrial Electronics (TIE) 66 (10) (2019) 8244–8252.
[71] Y. Sun, S. Wang, T. Hsieh, X. Tang, V. G. Honavar, MEGAN: A generative adversarial network for multi-view network embedding, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 3527–3533.
[72] M. Chen, L. Denoyer, Multi-view generative adversarial networks, in: Proceedings of the European Conference Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Vol. 10535, 2017, pp. 175–188.
[73] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Transactions on Neural Networks (TNN) 20 (1) (2009) 61–80.
[74] W. Huang, T. Zhang, Y. Rong, J. Huang, Adaptive sampling towards fast graph representation learning, in: Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS), 2018, pp. 4563–4572.
[75] W. L. Hamilton, P. Bajaj, M. Zitnik, D. Jurafsky, J. Leskovec, Embedding logical queries on knowledge graphs, in: Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS), 2018, pp. 2030–2041.
[76] K. Hassani, A. H. K. Ahmadi, Contrastive multi-view representation learning on graphs, in: Proceedings of the International Conference on Machine Learning (ICML), 2020, pp. 4116–4126.
[77] S. Fan, X. Wang, C. Shi, E. Lu, K. Lin, B. Wang, One2multi graph autoencoder for multi-view graph clustering, in: Proceedings of the International World Wide Web Conference (WWW), 2020, pp. 3070–3076.
[78] M. R. Khan, J. E. Blumenstock, Multi-gcn: Graph convolutional networks for multi-view networks, with applications to global poverty, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2019, pp. 606–613.
[79] H. Ma, Y. Bian, Y. Rong, W. Huang, T. Xu, W. Xie, G. Ye, J. Huang, Dual message passing neural network for molecular property prediction, CoRR abs/2005.13607.
[80] F. Xue, X. Wu, S. Cai, J. Wang, Learning multi-view camera relocalization with graph neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11372–11381.
[81] N. Zhang, S. Ding, J. Zhang, Y. Xue, An overview on restricted boltzmann machines, Neurocomputing 275 (2018) 1186–1199.
[82] M. R. Amer, T. J. Shields, B. Siddiquie, A. Tamrakar, A. Divakaran, S. M. Chai, Deep multimodal fusion: A hybrid approach, Internation Journal of Computer Vision (IJCV) 126 (2-4) (2018) 440–456.
[83] A. S. Al-Waisy, R. Qahwaji, S. S. Ipson, S. Al-Fahdawi, A multimodal deep learning framework using local feature representations for face recognition, Machince Vision and Applications (MVA) 29 (1) (2018) 35–54.
[84] A. F. Syafiandini, I. Wasito, S. Yazid, A. Fitriawan, M. Amien, Multimodal deep boltzmann machines for feature selection on gene expression

data, in: Proceedings of the International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2016, pp. 407–412.

[85] N. Zhang, S. Ding, H. Liao, W. Jia, Multimodal correlation deep belief networks for multi-view classification, Applied Intelligence 49 (5) (2019) 1925–1936.

[86] I. Sutskever, J. Martens, G. E. Hinton, Generating text with recurrent neural networks, in: Proceedings of the International Conference on Machine Learning (ICML), 2011, pp. 1017–1024.

[87] A. H. Abdulnabi, B. Shuai, Z. Zuo, L. Chau, G. Wang, Multimodal recurrent neural networks with information transfer layers for indoor scene labeling, IEEE Transanctions on Multimedia (TMM) 20 (7) (2018) 1656–1671.

[88] A. Sano, W. Chen, D. L. Martinez, S. Taylor, R. W. Picard, Multimodal ambulatory sleep detection using LSTM recurrent neural networks, IEEE Journal of Biomedical and Health Informatics (JBHI) 23 (4) (2019) 1607–1617.

[89] A. Narayanan, A. Siravuru, B. Dariush, Temporal multimodal fusion for driver behavior prediction tasks using gated recurrent fusion units, CoRR abs/1910.00628.
URL http://arxiv.org/abs/1910.00628

[90] S. Su, Z. Zhong, C. Zhang, Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 3027–3035.

[91] N. Tishby, N. Zaslavsky, Deep learning and the information bottleneck principle, in: Proceedings of the IEEE Information Theory Workshop (ITW), 2015, pp. 1–5.

[92] T. Kim, S. Wong, R. Cipolla, Tensor canonical correlation analysis for action classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[93] L. Sun, S. Ji, J. Ye, Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33 (1) (2011) 194–200.

[94] X. Z. Fern, C. E. Brodley, M. A. Friedl, Correlation clustering for learning mixtures of canonical correlation models, in: Proceedings of the SIAM International Conference on Data Mining (SDM), 2005, pp. 439–448.

[95] K. Chaudhuri, S. M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, in: Proceedings of the Annual International Conference on Machine Learning (ICML), 2009, pp. 129–136.

[96] T. Sun, S. Chen, Locality preserving CCA with applications to data visualization and pose estimation, Image and Vision Computing (IVC) 25 (5) (2007) 531–543.

[97] M. Kanai, R. Togo, T. Ogawa, M. Haseyama, Aesthetic quality assessment of images via supervised locality preserving CCA, in: Proceedings of the IEEE Global Conference on Consumer Electronics (GCCE), 2017, pp. 1–2.

[98] S. Becker, G. E. Hinton, Self-organizing neural network that discovers surfaces in random-dot stereograms, Nature 355 (6356) (1992) 161.

[99] S. Becker, Mutual information maximization: models of cortical self-organization, Network Computation in Neural Systems 7 (1) (1996) 7–31.

[100] W. W. Hsieh, Nonlinear canonical correlation analysis by neural networks, Neural Networks 13 (10) (2000) 1095–1105.

[101] A. Lu, W. Wang, M. Bansal, K. Gimpel, K. Livescu, Deep multilingual correlation for improved word embeddings, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 250–256.

[102] W. Wang, R. Arora, K. Livescu, N. Srebro, Stochastic optimization for deep CCA via nonlinear orthogonal iterations, in: Proceedings of the Annual Allerton Conference on Communication, Control, and Computing, 2015, pp. 688–695.

[103] F. Yan, K. Mikolajczyk, Deep correlation for matching images and text, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3441–3450.

[104] Q. Gao, H. Lian, Q. Wang, G. Sun, Cross-modal subspace clustering via deep canonical correlation analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 3938–3945.

[105] Z. Sun, P. K. Sarma, W. A. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 8992–8999.

[106] N. Tishby, F. Pereira, W. Bialek, The information bottleneck method, in: Proceedings of the Annual Allerton Conference on Communnication, Control and Computing, 1999, pp. 368–377.

[107] G. Chechik, A. Globerson, N. Tishby, Y. Weiss, Information bottleneck for gaussian variables, Journal of Machine Learning Research (JMLR) 6 (2005) 165–188.

[108] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, R. Arora, Deep generalized canonical correlation analysis, in: Proceedings of the Workshop on Representation Learning for NLP, 2019, pp. 1–6.

[109] N. Srebro, A. Shraibman, in: Proceedings of Annual International Conference on Learning Theory (ICLT), 2005, pp. 545–560.

[110] D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, in: Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS), 2000, pp. 556–562.

[111] C. H. Q. Ding, T. Li, M. I. Jordan, Convex and semi-nonnegative matrix factorizations, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 32 (1) (2010) 45–55.

[112] D. Cai, X. He, J. Han, T. S. Huang, Graph regularized nonnegative matrix factorization for data representation, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33 (8) (2011) 1548–1560.

[113] M. Zitnik, B. Zupan, Data fusion by matrix factorization, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 37 (1) (2015) 41–53.

[114] X. Cai, F. Nie, H. Huang, F. Kamangar, Heterogeneous image feature integration via multi-modal spectral clustering, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1977–1984.

[115] H. Huang, Y. Chuang, C. Chen, Affinity aggregation for spectral clustering, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 773–780.

[116] F. Nie, G. Cai, J. Li, X. Li, Auto-weighted multi-view learning for image clustering and semi-supervised classification, IEEE Transactions on Image Processing (TIP) 27 (3) (2018) 1501–1511.

[117] F. Nie, J. Li, X. Li, Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification, in: S. Kambhampati (Ed.), Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 1881–1887.

[118] J. Gao, J. Han, J. Liu, C. Wang, Multi-view clustering via joint nonnegative matrix factorization, in: Proceedings of SIAM International Conference on Data Mining (SDM), 2013, pp. 252–260.

[119] C. Hong, J. Yu, J. You, X. Chen, D. Tao, Multi-view ensemble manifold regularization for 3d object recognition, Information Sciences 320 (2015) 395–405.

[120] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, B. W. Schuller, A deep semi-nmf model for learning hidden representations, in: Proceedings of the International Conference on Machine Learning (ICML), Vol. 32, 2014, pp. 1692–1700.

[121] H. Zhao, Z. Ding, Y. Fu, Multi-view clustering via deep matrix factorization, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2017, pp. 2921–2927.

[122] C. Xu, Z. Guan, W. Zhao, Y. Niu, Q. Wang, Z. Wang, Deep multi-view concept learning, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 2898–2904.

[123] Z. Guan, L. Zhang, J. Peng, J. Fan, Multi-view concept learning for data representation, IEEE Transactions on Knowledge and Data Engineering (TKDE) 27 (11) (2015) 3016–3028.

[124] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 35 (1) (2013) 171–184.

[125] R. Mariappan, V. Rajan, Deep collective matrix factorization for augmented multi-view learning, Machine Learning 108 (8-9) (2019) 1395–1420.

[126] S. Huang, Z. Kang, Z. Xu, Auto-weighted multi-view clustering via deep matrix decomposition, Pattern Recognition (PR) 97 (2020) 1–11.

[127] S. Wei, J. Wang, G. Yu, C. Domeniconi, X. Zhang, Multi-view multiple clusterings using deep matrix factorization, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 6348–6355.

[128] U. von Luxburg, A tutorial on spectral clustering, Statistics and Computing 17 (4) (2007) 395–416.

[129] A. Kumar, P. Rai, H. D. III, Co-regularized multi-view spectral clustering, in: Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS), 2011, pp. 1413–1421.

[130] C. Lu, S. Yan, Z. Lin, Convex sparse spectral clustering: Single-view to multi-view, IEEE Transactions on Image Processing (TIP) 25 (6) (2016) 2833–2843.

[131] X. Yao, J. Han, D. Zhang, F. Nie, Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering, IEEE Transactions on Image Processing (TIP) 26 (7) (2017) 3196–3209.

[132] S. Zhou, X. Liu, J. Liu, X. Guo, Y. Zhao, E. Zhu, Y. Zhai, J. Yin, W. Gao, Multi-view spectral clustering with optimal neighborhood laplacian matrix, in: Proceeding of the AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 6965–6972.

[133] Y. Li, F. Nie, H. Huang, J. Huang, Large-scale multi-view spectral clustering via bipartite graph, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2015, pp. 2750–2756.

[134] M. T. Law, R. Urtasun, R. S. Zemel, Deep spectral clustering learning, in: Proceedings of the International Conference on Machine Learning (ICML), 2017, pp. 1985–1994.

[135] U. Shaham, K. P. Stanton, H. Li, R. Basri, B. Nadler, Y. Kluger, Spectralnet: Spectral clustering using deep neural networks, in: Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[136] X. Yang, C. Deng, F. Zheng, J. Yan, W. Liu, Deep spectral clustering using dual autoencoder network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4066–4075.

[137] Y. Wada, S. Miyamoto, T. Nakagama, L. Andéol, W. Kumagai, T. Kanamori, Spectral embedded deep clustering, Entropy 21 (8) (2019) 795.

[138] S. Affeldt, L. Labiod, M. Nadif, Spectral clustering via ensemble deep autoencoder learning (SC-EDAE), Pattern Recognition (PR) 108 (2020) 107522.

[139] X. Zhu, Y. Zhu, W. Zheng, Spectral rotation for deep one-step clustering, Pattern Recognition (PR) 105 (2020) 107175.

[140] G. Wen, Y. Zhu, W. Zheng, Spectral representation learning for one-step spectral rotation clustering, Neurocomputing 406 (2020) 361–370.

[141] Z. Hu, F. Nie, R. Wang, X. Li, Multi-view spectral clustering via integrating nonnegative embedding and spectral embedding, Information Fusion 55 (2020) 251–259.

[142] Z. Huang, J. T. Zhou, X. Peng, C. Zhang, H. Zhu, J. Lv, Multi-view spectral clustering network, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 2563–2569.

[143] S. Huang, K. Ota, M. Dong, F. Li, Multispectralnet: Spectral clustering using deep neural network for multi-view data, IEEE Transactions on Computational Social Systems (ITCSS) 6 (4) (2019) 749–760.

[144] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 1735–1742.

[145] Y. Gao, S. Gu, J. Li, Z. Liao, The multi-view information bottleneck clustering, in: Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA), 2007, pp. 912–917.

[146] X. Yan, Y. Ye, X. Qiu, M. Manic, H. Yu, CMIB: unsupervised image object categorization in multiple visual contexts, IEEE Transactions on Industrial Informatics (TII) 16 (6) (2020) 3974–3986.

[147] X. Yan, S. Hu, Y. Ye, Multi-task clustering of human actions by sharing information, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4049–4057.

[148] X. Yan, Z. Lou, S. Hu, Y. Ye, Multi-task information bottleneck co-clustering for unsupervised cross-view human action categorization, ACM Transactions on Knowledge Discovery from Data (ACM TKDD) 14 (2) (2020) 15:1–15:23.

[149] X. Yan, Y. Ye, Y. Mao, H. Yu, Shared-private information bottleneck method for cross-modal clustering, IEEE Access 7 (2019) 36045–36056.

[150] X. Yan, Y. Mao, S. Hu, Y. Ye, Heterogeneous dual-task clustering with visual-textual information, in: Proceedings of the SIAM International Conference on Data Mining (SDM), 2020, pp. 658–666.

[151] X. Yan, Y. Ye, X. Qiu, Unsupervised human action categorization with consensus information bottleneck method, in: Proceedings of the

International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 2245–2251.
[152] S. Hu, X. Yan, Y. Ye, Joint specific and correlated information exploration for multi-view action clustering, Information Sciences 524 (2020) 148–164.
[153] S. Hu, X. Yan, Y. Ye, Dynamic auto-weighted multi-view co-clustering, Pattern Recognition (PR) 99 (2020) 1–12.
[154] A. A. Alemi, I. Fischer, J. V. Dillon, K. Murphy, Deep variational information bottleneck, in: Proceedings of the International Conference on Learning Representations (ICLR), 2017.
[155] T. T. Nguyen, J. Choi, Markov information bottleneck to improve information flow in stochastic neural networks, Entropy 21 (10) (2019) 976.
[156] A. A. Alemi, Variational predictive information bottleneck, in: Symposium on Advances in Approximate Bayesian Inference (AABI), Vol. 118, 2019, pp. 1–6.
[157] Q. Wang, C. Boudreau, Q. Luo, P. Tan, J. Zhou, Deep multi-view information bottleneck, in: Proceedings of the SIAM International Conference on Data Mining (SDM), 2019, pp. 37–45.
[158] M. Federici, A. Dutta, P. Forré, N. Kushman, Z. Akata, Learning robust representations via multi-view information bottleneck, in: Proceedings of the International Conference on Learning Representations (ICLR), 2020.
[159] G. Wu, Z. Lin, J. Han, L. Liu, G. Ding, B. Zhang, J. Shen, Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 2854–2860.
[160] C. Li, C. Deng, N. Li, W. Liu, X. Gao, D. Tao, Self-supervised adversarial hashing networks for cross-modal retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4242–4251.
[161] C. Li, C. Deng, L. Wang, D. Xie, X. Liu, Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2019, pp. 176–183.
[162] R. Xu, C. Li, J. Yan, C. Deng, X. Liu, Graph convolutional network hashing for cross-modal retrieval, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 982–988.
[163] C. Deng, Z. Chen, X. Liu, X. Gao, D. Tao, Triplet-based deep hashing network for cross-modal retrieval, IEEE Transactions on Image Processing (TIP) 27 (8) (2018) 3893–3903.
[164] L. Wu, Y. Wang, L. Shao, Cycle-consistent deep generative hashing for cross-modal retrieval, IEEE Transactions on Image Processing (TIP) 28 (4) (2019) 1602–1612.
[165] Y. Shi, X. You, F. Zheng, S. Wang, Q. Peng, Equally-guided discriminative hashing for cross-modal retrieval, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 4767–4773.
[166] L. Zhen, P. Hu, X. Wang, D. Peng, Deep supervised cross-modal retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10394–10403.
[167] P. Huang, X. Chang, A. G. Hauptmann, E. H. Hovy, Forward and backward multimodal NMT for improved monolingual and multilingual cross-modal retrieval, in: Proceedings of the International Conference on Multimedia Retrieval (ICMR), 2020, pp. 53–62.
[168] P. Dou, I. A. Kakadiaris, Multi-view 3d face reconstruction with deep recurrent neural networks, Image Vision and Computation (IVC) 80 (2018) 80–91.
[169] S. Bi, Z. Xu, K. Sunkavalli, D. J. Kriegman, R. Ramamoorthi, Deep 3d capture: Geometry and reflectance from sparse multi-view images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5959–5968.
[170] S. S. Farfade, M. J. Saberian, L. Li, Multi-view face detection using deep convolutional neural networks, in: Proceedings of the ACM on International Conference on Multimedia Retrieval (ICMR), 2015, pp. 643–650.
[171] Z. Bai, Z. Cui, J. A. Rahim, X. Liu, P. Tan, Deep facial non-rigid multi-view stereo, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5849–5859.
[172] J. Li, J. Zhao, F. Zhao, H. Liu, J. Li, S. Shen, J. Feng, T. Sim, Robust face recognition with deep multi-view representation learning, in: Proceedings of the ACM Conference on Multimedia Conference (ACM'MM), 2016, pp. 1068–1072.
[173] Y. Guo, Y. Xia, J. Wang, H. Yu, R. Chen, Real-Time Facial Affective Computing on Mobile Devices. Sensors (Basel, Switzerland) vol. 20,3 870. 6 Feb. 2020,
[174] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, K. Yan, A deep neural network-driven feature learning method for multi-view facial expression recognition, IEEE Transactions on Multimedia (TMM) 18 (12) (2016) 2528–2536.
[175] Y. Xia, H. Yu, F. Wang, Accurate and robust eye center localization via fully convolutional networks, IEEE/CAA Journal of Automatica Sinica 6 (5) (2019) 1127–1138.
[176] F. Zhao, J. Li, L. Zhang, Z. Li, S. Na, Multi-view face recognition using deep neural networks, Future Generation Computer Systems 111 (2020) 375–380.
[177] Y. Wang, X. Dong, G. Li, J. Dong, H. Yu, Cascade regression-based face frontalization for dynamic facial expression analysis, Cognitive Computation, 2021.
[178] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, Skeleton-indexed deep multi-modal feature learning for high performance human action recognition, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2018, pp. 1–6.
[179] A. A. Alani, G. Cosma, A. Taherkhani, Classifying imbalanced multi-modal sensor data for human activity recognition in a smart home using deep learning, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–8.
[180] M. Trumble, A. Gilbert, A. Hilton, J. P. Collomosse, Deep convolutional networks for marker-less human pose estimation from multiple views, in: Proceedings of the European Conference on Visual Media Production (CVMP), 2016, pp. 6:1–6:9.
[181] F. Huang, A. Zeng, M. Liu, Q. Lai, Q. Xu, Deepfuse: An imu-aware network for real-time 3d human pose estimation from multi-view image, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 418–427.
[182] X. Zhang, D. Ma, H. Yu, Y. Huang, P. Howell, B. Stevens, Scene perception guided crowd anomaly detection, Neurocomputing 414 (2020) 291–302.
[183] D. Tao, Y. Guo, B. Yu, J. Pang, Z. Yu, Deep multi-view feature learning for person re-identification, IEEE Transactions on Circuits and Systems for Video Technology (ITCSVT) 28 (10) (2018) 2657–2666.
[184] Y. Zhou, L. Liu, L. Shao, Vehicle re-identification by deep hidden multi-view inference, IEEE Transactions on Image Processing (TIP)

27 (7) (2018) 3275–3287.
[185] X. Xin, X. Wu, Y. Wang, J. Wang, Deep self-paced learning for semi-supervised person re-identification using multi-view self-paced clustering, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2019, pp. 2631–2635.
[186] W. Liu, Z. Wang, X. Liu, N. Zeng, D. Bell, A novel particle swarm optimization approach for patient clustering from emergency departments, IEEE Transactions on Evolutionary Computation (TEC) 23 (4) (2019) 632–644.
[187] N. Zeng, H. Li, Z. Wang, W. Liu, X. Liu, Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip, Neurocomputing (2020) 1–8.
[188] W. Yue, Z. Wang, W. Liu, B. Tian, S. Lauria, X. Liu, An optimally weighted user- and item-based collaborative filtering approach to predicting baseline data for friedreich's ataxia patients, Neurocomputing 419 (2021) 287–294.
[189] N. Zeng, Z. Wang, H. Zhang, K. Kim, Y. Li, X. Liu, An improved particle filter with a novel hybrid proposal distribution for quantitative analysis of gold immunochromatographic strips, IEEE Transactions on Nanotechnology (TN) 18 (1) (2019) 819–829.
[190] X. Fei, L. Shen, S. Ying, Y. Cai, Q. Zhang, W. Kong, W. Zhou, J. Shi, Parameter transfer deep neural network for single-modal b-mode ultrasound-based computer-aided diagnosis, Cognitive Computation 12 (6) (2020) 1252–1264.
[191] D. H. Kim, S. Kim, Y. M. Ro, Latent feature representation with 3-d multi-view deep convolutional neural network for bilateral analysis in digital breast tomosynthesis, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 927–931.
[192] P. Jonnalagedda, D. Schmolze, B. Bhanu, Mvpnets: Multi-viewing path deep learning neural networks for magnification invariant diagnosis in breast cancer, in: Proceedings of the IEEE International Conference on Bioinformatics and Bioengineering (BIBE), 2018, pp. 189–194.
[193] B. Gong, L. Shen, C. Chang, S. Zhou, W. Zhou, S. Li, J. Shi, Bi-modal ultrasound breast cancer diagnosis via multi-view deep neural network SVM, in: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1106–1110.
[194] Y. Yuan, G. Xun, F. Ma, Q. Suo, H. Xue, K. Jia, A. Zhang, A novel channel-aware attention framework for multi-channel EEG seizure detection via multi-view deep learning, in: Proceedings of the IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), 2018, pp. 206–209.
[195] Y. Yuan, G. Xun, K. Jia, A. Zhang, A multi-view deep learning framework for EEG seizure detection, IEEE Journal of Biomedical and Health Informatics 23 (1) (2019) 83–94.
[196] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, J. A. Noble, Ω-net (omega-net): Fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks, Medical Image Analysis (MIA) 48 (2018) 95–106.
[197] Y. Liu, M. Yin, S. Sun, Multi-view learning and deep learning for microscopic neuroblastoma pathology image diagnosis, in: Proceedings of the Pacific Rim International Conference on Artificial Intelligence (PRICAI), 2018, pp. 545–558.
[198] Y. Pi, Z. Zhao, Y. Xiang, Y. Li, H. Cai, Z. Yi, Automated diagnosis of bone metastasis based on multi-view bone scans using attention-augmented deep neural networks, Medical Image Analysis (MIA) 65 (2020) 101784.
[199] F. Li, R. Fergus, P. Perona, Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories, Computer Vision and Image Understanding (CVIU) 106 (1) (2007) 59–70.
[200] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y.-T. Zheng, Nus-wide: A real-world web image database from national university of singapore, in: Proceedings of the ACM Conference on Image and Video Retrieval (CIVR'09), 2009.
[201] M. Nilsback, A. Zisserman, A visual vocabulary for flower classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 1447–1454.
[202] F. Li, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 524–531.
[203] M. Amini, N. Usunier, C. Goutte, Learning from multiple partially observed views - an application to multilingual text categorization, in: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), 2009, pp. 28–36.
[204] S. Ramagiri, R. Kavi, V. Kulathumani, Real-time multi-view human action recognition using a wireless camera network, in: R. P. Kleihorst, A. Prati, S. Velipasalar (Eds.), Proceedings of the ACM/IEEE International Conference on Distributed Smart Cameras, 2011, pp. 1–6.
[205] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3d exemplars, in: Proceedings of the IEEE International Conference on Computer Vision (CVPR), 2007, pp. 1–7.
[206] A. Liu, N. Xu, W. Nie, Y. Su, Y. Wong, M. Kankanhalli, Benchmarking a multimodal and multiview and interactive dataset for human action recognition, IEEE Transactions on Cybernetics (TCYB) 47 (7) (2017) 1781–1794.
[207] D. Zeng, Y. Yu, K. Oyama, Audio-visual embedding for cross-modal music video retrieval through supervised deep CCA, in: Proceedings of the IEEE International Symposium on Multimedia (ISM), 2018, pp. 143–150.
[208] Y. Zhou, Z. Wang, C. Fang, T. Bui, T. L. Berg, Visual to sound: Generating natural sound for videos in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3550–3558.
[209] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1290–1297.
[210] J. Wang, X. Nie, Y. Xia, Y. Wu, S. Zhu, Cross-view action modeling, learning, and recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2649–2656.
[211] A. Liu, Y. Su, P. Jia, Z. Gao, T. Hao, Z. Yang, Multipe/single-view human action recognition via part-induced multitask structural learning, IEEE Transactions on Cybernetics (TCYB) 45 (6) (2015) 1194–1208.