Deep Supervised Multi-View Learning with Graph Priors

Peng Hu, Liangli Zhen, Xi Peng, Hongyuan Zhu, Jie Lin, Xu Wang, Dezhong Peng

***Abstract*—This paper presents a novel method for supervised multi-view representation learning, which projects multiple views into a latent common space while preserving the discrimination and intrinsic structure of each view. Specifically, an *apriori* discriminant similarity graph is first constructed based on labels and pairwise relationships of multi-view inputs. Then, view-specific networks progressively map inputs to common representations whose affinity approximates the constructed graph. To achieve graph consistency, discrimination, and cross-view invariance, the similarity graph is enforced to meet the following constraints: 1) pairwise relationship should be consistent between the input space and common space for each view; 2) within-class similarity is larger than any between-class similarity for each view; 3) the inter-view samples from the same (or different) classes are mutually similar (or dissimilar). Consequently, the intrinsic structure and discrimination are preserved in the latent common space using an *apriori* approximation schema. Moreover, we present a sampling strategy to approach a sub-graph sampled from the whole similarity structure instead of approximating the graph of the whole dataset explicitly, thus benefiting lower space complexity and the capability of handling large-scale multi-view datasets. Extensive experiments show the promising performance of our method on five datasets by comparing it with 18 state-of-the-art methods.**

***Index Terms*—Structure preservation, discriminant structure, common space, cross-view recognition, cross-modal retrieval.**

I. Introduction

As multi-view data, such as images, textual descriptions, and videos, continues to rapidly grow, there is an increasing demand for developing multi-view learning approaches to cater to a wide range of applications, including multimedia retrieval [1]–[4], image annotation [5], heterogeneous face recognition [6], and cross-view retrieval [7], [8]. It is a fundamental challenge in multi-view learning to measure the similarity between samples from different views, commonly known as the '*heterogeneous gap*' [1], [9], [10].

To solve this fundamental problem, a number of multi-view methods [2], [6], [11]–[13] have been developed by projecting data samples from distinct views into a shared space. The traditional ones [14]–[16] adopt the statistical correlation analysis technique to find the projections by maximizing the correlations between different views without category annotations. To exploit the label information, a variety of multi-view methods have been developed in a supervised manner [10], [11], [17], [18]. In recent years, the graph regularization technique has been employed to uncover the intrinsic structure of the dataset to boost the performance of the semi-supervised and supervised multi-view learning methods [19]–[21], and they have achieved promising results. Despite their potential, these graph regularization-based multi-view methods have not sufficiently leveraged label information, and many of them are designed to learn only linear projections, which restricts their ability to handle the complex, nonlinear nature of many real-world applications. Additionally, these methods are limited by their requirement to compute the similarity graph using the entire training dataset, resulting in high computational and space complexity, which impedes their efficiency and effectiveness in dealing with large-scale multi-view datasets.

To overcome the two problems, in this paper, we propose a novel deep multi-view learning method using *apriori* approximation (DMLPA) to learn nonlinear transformations for multi-view recognition and retrieval. It approximates *apriori* similarity graph which is constructed in advance based on the samples in the input space to preserve the intrinsic structure and the discrimination in the latent common space using the *apriori* approximation schema. More specifically, DMLPA consists of v view-specific networks and a novel graph-based loss as shown in Figure 1. The similarity graph matrix is computed based on the pairwise distances of intra-view samples and their category labels, which aims to sample a sub-graph from the global graph structure. To uncover the intrinsic structure of the dataset, the following three constraints are enforced and formulated in our *apriori* similarity graph: 1) in the same view, the similarity of nearby samples from the same (resp. different) classes is larger than that of farther samples from the same (resp. different) classes, thus preserving the graphic information into the similarity graph for every single view; 2) the within-class similarity is larger than any between-class similarity, leading to preserving the discrimination into the similarity graph; 3) the samples of different views from the same (resp. different) classes should be sufficiently similar

This work was supported by the National Natural Science Foundation of China (U19A2078, 62102274, 62372315, 61971296, and 62306197), Sichuan Science and Technology Planning Project (2023YFG0033, 2023ZHCG0016, 2023YFQ0020, and 23ZYZYTS0077), Chengdu Science and Technology Project (2023-XT00-00004-GX and 2023-GH02-00064-HZ), the Fundamental Research Funds for the Central Universities under Grant YJ202140, and the National Research Foundation of Singapore under its AI Singapore Programme (AISG Award No.: AISG2-TC-2021-003).

P. Hu is with the College of Computer Science, Sichuan University, Chengdu 610065, China, and also with the State Key Laboratory of Integrated Service Networks (Xidian University), Xi'an 710071, China.

L. Zhen is with the Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138632.

X. Peng and X. Wang are with the College of Computer Science, Sichuan University, Chengdu 610065, China.

H. Zhu and J. Lin are with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632.

D. Peng is with the College of Computer Science, Sichuan University, Chengdu 610065, China, and also with Chengdu Ruibei Yingte Information Technology Co., Ltd, Chengdu 610054, China

Corresponding author: Liangli Zhen (email: zhenll@ihpc.a-star.edu.sg).

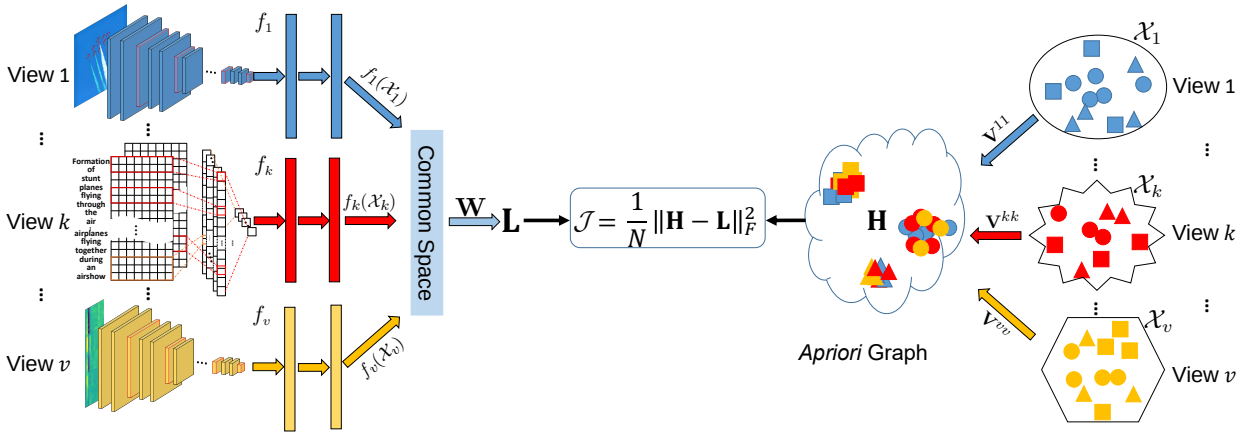


Fig. 1: The framework of DMLPA. In the figure, distinct shapes are used to represent diverse classes and distinct colors are used to denote different views. $\mathbf{W}$ and $\mathbf{V}^{kk}$ are the similarity matrices of all common representations and the k-th view inputs $\mathcal{X}_k$, respectively. $\mathbf{L}$ and $\mathbf{H}$ are the normalized graph Laplacian matrices that represent the graphs of common space and input data, respectively. Moreover, $\mathbf{L}$ and $\mathbf{H}$ are respectively computed by $\mathbf{W}$ and $\mathbf{V}^{kl}|_{k,l}^{v}$ (see Equation (8) and Equation (11)), where $\mathbf{V}^{kl}|_{k\neq l}^{v}$ are inter-view similarity matrices computed by intra-view similarity matrices $\mathbf{V}^{kk}|_{k}^{v}$ and labels (see Equation (5)). $\mathcal{J} = \frac{1}{N}\|\mathbf{H}-\mathbf{L}\|_F^2$ is the loss to make the obtained common representations approximate *apriori* similarity graph of input data.

(resp. dissimilar), thus eliminating the view discrepancy across different views.

There are several key differences between our DMLPA and the existing multi-view methods [2], [10], [20], [22]–[24]. Firstly, different from [6], [23], [25], [26] which cannot preserve the intrinsic structure of multi-view inputs in the common space, our DMLPA preserves not only the graph information but also the discrimination of the multi-view data. Secondly, compared with traditional supervised graph regularization methods [11], [21], [22], [24] which utilize a constructed graph as a regularizer for their framework on all training data, DMLPA directly approximates *apriori* graph computed from labels and pairwise distances with a batch-by-batch schema. In consequence, our DMLPA exhibits significantly lower time and space complexity compared to existing graph regularization-based methods that require computing the graph for the entire training dataset. Furthermore, most of these methods [19]–[21], [23] are linear multi-view methods, whereas our DMLPA is a deep learning-based method that is capable of learning nonlinear transformations from more than two views. At last, comparing with methods presented in [27]–[30], which learn low-dimensional embeddings by minimizing the Kullback-Leibler (KL) divergence of two probability distributions within a single view, our method provides a unified model for handling multi-view data.

The framework of DMLPA is presented in Figure 1, which illustrates that samples from different views are projected into a latent common space by multiple view-specific networks to approximate an *apriori* similarity graph. This process enables the transfer of discrimination and intrinsic structure from the similarity graph to the common representations. In summary, the primary contributions of this work can be summarized as follows:

- This study proposes a novel deep multi-view learning method that leverages *apriori* approximation to project multi-view data into a common discriminant space. The proposed method utilizes label information to minimize discrimination errors and construct *apriori* similarity graphs that enforce graph consistency for enhanced multi-view learning. In other words, this approach allows for more effective representations of multi-view data by taking into account both label information and graph consistency.
- The *apriori* similarity graph is constructed to encapsulate the discrimination and intrinsic structure of the multi-view inputs. Since the labels and the pairwise relationships are simultaneously exploited to construct the similarity graph, the discrimination and pairwise information can be well preserved in the graph. Then, this graph is used to guide the networks to embed representations into a common space. In this way, as much discrimination as possible is preserved in the common representations, so does the intrinsic structure.
- Instead of using the graphic information as a regularizer, we directly sample a multi-view sub-graph from the global graph structure and make the multiple networks approximate the graph in a batch-by-batch manner. Thus, our model can be trained with a batch-by-batch schema and costs much less computation and storage resources than most of the existing cross-view graph regularization methods that have to calculate the separate graph for each view from the entire training dataset.

II. Related Work

In recent decades, numerous approaches for multi-view representation learning have been proposed to facilitate the learning of common representations to correlate heterogeneous data. A notable pioneer in this field is canonical correlation analysis (CCA) [31], which seeks to learn two transformations that can map distinct views into a shared space by maximizing dual-view correlation. Nevertheless, CCA is restricted to handling dual-view data, which gave rise to the

development of multi-view CCA (MCCA) [14], [32]. These methods endeavor to learn v view-specific transformations for v views by maximizing the correlations across diverse views. Another common unsupervised dual-view approach is partial least squares (PLS) [15] that linearly maps distinct views into a shared space while maximizing the covariance of the two views. This ensures that distinct views are highly correlated in the latent shared space. Additionally, Kan et al. proposed multi-view discriminant analysis (MvDA) and MvDA with view-consistency (MvDA-VC) methods that employ Fisher's criterion to learn multiple view-specific linear transformations, enabling them to map distinct views into a latent shared space [17]. Furthermore, these linear approaches can be extended to nonlinear versions using the kernel trick, *e.g.* kernel canonical correlation analysis [16], [33], [34]. However, the learned representations will be limited by the predetermined kernels.

In addition, Deep Neural Networks (DNNs) with strong nonlinear correlation modeling capability have made significant progress in various single-view tasks, such as image retrieval, and subspace clustering [35], [36]. To endow multi-view learning methods with the capability, DNNs have been utilized to model cross-view correlations [12], [13], [18], [37], [38]. To be specific, unsupervised works attempted to learn complex nonlinear transformations of dual-view data to get highly linearly correlated representations [37]–[39]. Meanwhile, supervised methods have explored both intra- and inter-view correlations to learn a common space [2], [10], [25]. With the success of generative adversarial nets (GANs), some works aim to seek an effective common discriminant space based on adversarial learning [13], [40], [41]. In [42], the authors presented a modality-specific cross-modal similarity measurement (MCSM) method to directly estimate cross-view similarity without an explicit common representation. However, few of these methods are specifically designed to handle more than two views while preserving the view-specific intrinsic structures, resulting in being limited to two views only and ignoring some beneficial graphic information in the multiple views.

Recently, inspired by the great success of graph theory applications in many other fields [36], [43], [44], some methods are proposed to learn view-specific linear transformations by preserving the intrinsic geometric structures of original views [11], [21], [45], [46]. Cross-view graph regularization is used to enrich the multi-view training data and makes the solution smooth. It has achieved promising results in the cross-view retrieval problem by preserving the intra- and inter-view similarity relationships [19]–[21]. Despite the effectiveness, cross-view graph regularization-based methods have a drawback in that they calculate the graph based on the entire training dataset, resulting in prohibitively high time and space complexity for handling large-scale multi-view datasets. Furthermore, the majority of these methods are linear, rendering them unsuitable for handling the high levels of nonlinear complexity present in many real-world applications.

III. Our Proposed Method

Let $\mathcal{X}_k = \{\mathbf{X}_i^k \in \mathbb{R}^{d_k \times p_k} \mid i = 1, \cdots, N_k\}$ be the points of the k-th view, where $\mathbf{X}_i^k$ is the i-th point of the k-th view with

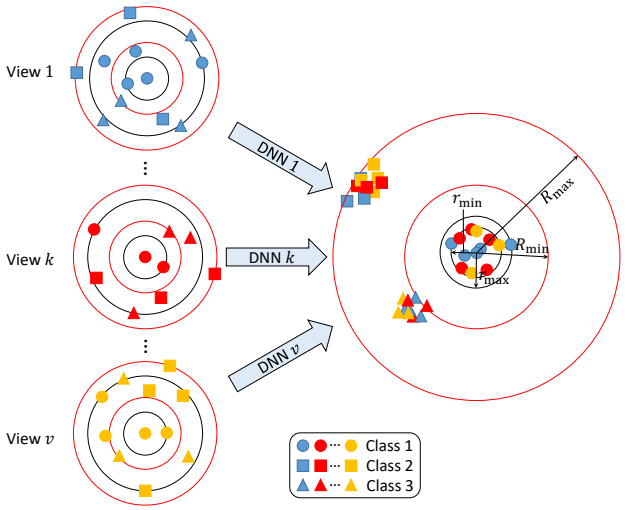


Fig. 2: The basic idea of our DMLPA. In the figure, distinct shapes represent diverse classes, and distinct colors denote different views. $r_{\min}$ and $r_{\max}$ respectively represent the minimum and maximum distances between the center point and the corresponding within-class neighbors (black circles). $R_{\min}$ and $R_{\max}$ respectively represent the minimum and maximum distances between the center point and the corresponding between-class samples (red circles). It aims at preserving the discrimination and intrinsic structure into an *apriori* graph.

the dimensionality of $d_k \times p_k$, and N_k is the number of points from the k-th view. The view-specific network of the k-th view could be denoted as a nonlinear function $f_k(\cdot; \Theta_k) \in \mathbb{R}^{q \times 1}$ as shown in Figure 1, where q is the objective dimensionality of the shared space and Θ_k are the network parameters. Therefore, the output of the k-th view could be formulated as

$$\mathbf{y}_i^k = f_k(\mathbf{X}_i^k) \tag{1}$$

for the sample $\mathbf{X}_i^k$. Moreover, the heat kernel is used to estimate the similarity between any two points in the common space as follows:

$$W_{ij}^{kl} = e^{-\frac{\left(d(\mathbf{y}_i^k, \mathbf{y}_j^l)\right)^2}{\tau}}, \tag{2}$$

where $d(\mathbf{y}_i^k, \mathbf{y}_j^l)$ is the distance between $\mathbf{y}_i^k$ and $\mathbf{y}_j^l$ and τ is a temperature parameter. From the formulation, we can see that the compact samples have a larger similarity than the scattered points. The similarity W_{ij}^{kl} can compose a similarity matrix $\mathbf{W}$ to present the geometric structure of the corresponding common representations, and it is desirable that the obtained common representations approximate the *apriori* graph of the input data.

Graph theory suggests that two samples are deemed similar if there is a high edge weight between the two points [21], [47]. Therefore, it is imperative to preserve the intrinsic structure specific to each view, enhance the similarity within the same class, decrease the similarity between different classes, and eliminate any view-related discrepancies in the shared space. To achieve these goals, there are three following objectives:

- It is desirable to preserve the intrinsic structure of each view in the learned space, *i.e.* for each view, the similarity

of the samples can be consistent between the common space and the view-specific input spaces.

- It is appropriate to make the same classes compact while the distinct classes scatter for each view in the shared space, *i.e.* the within-class similarity is larger than the between-class one in each view. Therefore, discrimination can be preserved for each view in the shared space.
- It is expected to minimize the discrepancy across distinct views in the shared space. In other words, the inter-view similarity of the same (resp. different) class can be sufficiently larger (resp. smaller) than the between-class (resp. within-class) similarity in the intrinsic graph. Thus, this leads to a compact arrangement of data points belonging to the same category, and a scattered distribution of points belonging to diverse categories in the common space even if they originate from different views.

The basic idea of our DMLPA is depicted in Figure 2. Similar to Equation (2), we define the similarity between input data samples using the following equation:

$$V_{ij}^{kl} = e^{-\frac{\left(d(\mathbf{x}_i^k, \mathbf{x}_j^l)\right)^2}{t}}. \tag{3}$$

To achieve the first objective, the *apriori* graph of the k-th view can be described by the matrix $\mathbf{V}^{kk}$, which is equivalent to sampling a sub-graph from the global graph. Moreover, to achieve the second objective, which is to enlarge the similarity of the same class and reduce the similarity of the distinct classes, label information is introduced to compute the intra-view graph matrix $\mathbf{V}^{kk}$ for k-th view as follows:

$$V_{ij}^{kk} = \begin{cases} e^{-\frac{\left(d(\mathbf{x}_i^k, \mathbf{x}_j^k)\right)^2}{t_1}}, & \text{if } \ell(\mathbf{X}_i^k) = \ell(\mathbf{X}_j^k); \\ e^{-\frac{\left(d(\mathbf{x}_i^k, \mathbf{x}_j^k)\right)^2}{t_2}}, & \text{otherwise,} \end{cases} \tag{4}$$

where $\ell(\cdot)$ is a function to obtain the class label of the corresponding sample, $t_1 = \alpha_1 \max\{(d(\mathbf{X}_i^k, \mathbf{X}_j^k))^2 | \ell(\mathbf{X}_i^k) = \ell(\mathbf{X}_j^k); i, j = 1, 2, \cdots, N_k\}$, $\max\{\cdot, \cdot\}$ gets the maximum value, $t_2 = \alpha_2 \min\{(d(\mathbf{X}_i^k, \mathbf{X}_j^k))^2 | \ell(\mathbf{X}_i^k) \neq \ell(\mathbf{X}_j^k); i, j = 1, 2, \cdots, N_k\}$, $\min\{\cdot, \cdot\}$ gets the minimum value, α_1 and α_2 are positive balance parameters with $0 < \alpha_2 < \alpha_1$. Therefore, it can be ensured that the within-class similarity is larger than every between-class similarity for each view. However, since the spaces of different views may be different and the graph of each view is absolutely disparate, the inter-view similarity matrix $V_{ij}^{kl}|_{k \neq l}$ cannot be directly computed by $\mathbf{X}_i^k$ and $\mathbf{X}_j^l$. To achieve the third objective, the within-class (resp. between-class) similarity in different views is desired to be as large (resp. small) as the one in each view. Then, we set the within-class similarity in different views as the largest one for the same class in all individual views. Similarly, the between-class similarity in different views is set as the smallest one for the different classes in all individual views. Thus, the inter-view samples from the same category will be compacted while the inter-view ones from the distinct classes will be scattered. The inter-view graph matrix can be formulated as:

$$V_{ij}^{kl} = \begin{cases} e^{-\frac{\left(\min\left\{r(\mathbf{x}_i^k), r(\mathbf{x}_j^l)\right\}\right)^2}{t_1}}, & \text{if } \ell(\mathbf{X}_i^k) = \ell(\mathbf{X}_j^l); \\ e^{-\frac{\left(\max\left\{R(\mathbf{x}_i^k), R(\mathbf{x}_j^l)\right\}\right)^2}{t_2}}, & \text{otherwise,} \end{cases} \tag{5}$$

where $r(\cdot)$ calculates the minimum distance between the corresponding point and its within-class samples; $R(\cdot)$ calculates the maximum distance between the corresponding point and its between-class samples. It guarantees that the within-class similarity of distinct views is larger than the between-class similarity of the distinct views, *i.e.*, it is to make the samples of the same classes from distinct views compact and the samples of the diverse classes from distinct views scatter.

Finally, the *apriori* graph of the points in a shared space can be described by a partitioned matrix as

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}^{11} & \cdots & \mathbf{W}^{1v} \\ \vdots & \ddots & \vdots \\ \mathbf{W}^{v1} & \cdots & \mathbf{W}^{vv} \end{bmatrix} \tag{6}$$

with the (k, l)-th submatrix

$$\mathbf{W}^{kl} = \begin{bmatrix} W_{11}^{kl} & \cdots & W_{1N_l}^{kl} \\ \vdots & \ddots & \vdots \\ W_{N_k1}^{kl} & \cdots & W_{N_kN_l}^{kl} \end{bmatrix}, \tag{7}$$

where each item W_{ij}^{kl} of $\mathbf{W}^{kl}$ can be computed by Equation (2). To weaken the influence of the data distribution differences on the global similarity measurement, the normalized graph Laplacian matrix is used to represent the graph of all views. Let $\mathbf{D}$ be the diagonal matrix with $D_{ii} = \sum_{j=1}^{N} W_{ij}$, where $N = \sum_{k=1}^{v} N_k$ is the number of all points in all views. From [48], we could formulate the the normalized *apriori* graph matrix $\mathbf{L}$ as follows:

$$\begin{aligned} \mathbf{L} &= \mathbf{D}^{-1}(\mathbf{D} - \mathbf{W}) \\ &= \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}. \end{aligned} \tag{8}$$

Similarly, the similarity graph also can be formulated as a partitioned matrix:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}^{11} & \cdots & \mathbf{V}^{1v} \\ \vdots & \ddots & \vdots \\ \mathbf{V}^{v1} & \cdots & \mathbf{V}^{vv} \end{bmatrix} \tag{9}$$

with the (k, l)-th submatrix

$$\mathbf{V}^{kl} = \begin{bmatrix} V_{11}^{kl} & \cdots & V_{1N_l}^{kl} \\ \vdots & \ddots & \vdots \\ V_{N_k1}^{kl} & \cdots & V_{N_kN_l}^{kl}, \end{bmatrix} \tag{10}$$

where each item V_{ij}^{kl} of $\mathbf{V}^{kl}$ can be obtained from Equation (4) and Equation (5). Let $\mathbf{E}$ be a diagonal matrix with $E_{ii} = \sum_{j=1}^{N} V_{ij}$. Similar to Equation (8), we could formulate the normalized Laplacian matrix $\mathbf{H}$ as:

$$\begin{aligned} \mathbf{H} &= \mathbf{E}^{-1}(\mathbf{E} - \mathbf{V}) \\ &= \mathbf{I} - \mathbf{E}^{-1}\mathbf{V}. \end{aligned} \tag{11}$$

Then, $\mathbf{L}$ and $\mathbf{H}$ can be used to present the intrinsic structures of the common representations and the multi-view data, respectively. To transfer the discrimination and intrinsic structure in the multiple input views to the common representations, the graph $\mathbf{L}$ of the common space is desired to approximate the

apriori graph $\mathbf{H}$ of the inputs. A novel objective function could be formulated as:

$$\begin{aligned}\mathcal{J} &= \frac{1}{N}\|\mathbf{H}-\mathbf{L}\|_F^2 \\ &= \frac{1}{N}\|\mathbf{D}^{-1}\mathbf{W}-\mathbf{E}^{-1}\mathbf{V}\|_F^2\end{aligned} \tag{12}$$

where $\|\cdot\|_F$ is Frobenius norm. The objective function enables the optimizer to train the multi-view networks in a batch-by-batch manner using back-propagation. The optimization process is further elaborated in Algorithm 1.

Algorithm 1 Optimization procedure of DMLPA

Input: The training data $\mathcal{X}_k|_{k=1}^v$, objective dimensionality q, batch size N_b, positive balance parameters α_1, α_2 and τ, learning rate β

1: **while** not converge **do**
2: Randomly select N_b samples for each view from $\mathcal{X}^k|_{k=1}^v$ to construct a multi-view mini-batch.
3: Compute the normalized *apriori* similarity graph matrix $\mathbf{H}$ according to Equations (4), (5) and (9) to (11) on a mini-batch.
4: Compute the common representations $\mathbf{y}^k|_{k=1}^v$ by the corresponding view-specific networks $f_k|_{k=1}^v$ for all views according to Equation (1).
5: Compute the similarity graph matrix $\mathbf{L}$ of the common representations according to Equations (2) and (6) to (8).
6: Update the parameters of view-specific networks $\Theta_k|_{k=1}^v$ by minimizing $\mathcal{J}$ in Equation (12) with descending their stochastic gradient: $\Theta_k = \Theta_k - \beta\frac{\partial\mathcal{J}}{\partial\Theta_k}$ $(k=1,\cdots,v)$
7: **end while**

Output: The optimized DMLPA model.

IV. Experiments

To assess the effectiveness of the proposed methods, we conducted experiments on five datasets: Reuters [49], [50], noisy MNIST [37], [51], Spoken Arabic Digit [52], [53] (nM-SAD), Pascal Sentence [54], XMediaNet [42], and MS-COCO [55].

A. Experiment Settings

1) Datasets and Features: Table I provides a summary of the statistics for the five widely-used benchmark datasets. For Reuters, the feature dimensions are greater than 10,000, making it difficult for most methods to handle. To address this problem, principal component analysis (PCA) is utilized to reduce the dimensionality of high-dimensional features to 500 dimensions. The nMSAD dataset is generated by combining the noisy MNIST [37], [51] and Spoken Arabic Digit datasets (SAD) [52], [53], which has two image views and one audio view. MV1 and MV2 are respectively two image views in noisy MNIST. Each image of the two image views is presented by a 28×28 grayscale digit matrix and each sample of the audio view is a 25×13 MFCC matrix.

Furthermore, to ensure a fair comparison, the all datasets were randomly divided into training, validation, and test subsets as illustrated in Table I. For the Pascal Sentence and XMediaNet datasets, we follow the dataset partitions of [42], [56]. It is worth noting that all the baselines employed the same image and text features, which were extracted from the CNN and Word2Vec models in our experiments. Specifically, we obtained the CNN feature of an image from the fc7 layer in the 19-layer VGGNet [57] pre-trained on the ImageNet. For the text modality, the pre-trained Word2Vec model [58], which was trained on billions of words in Google News, is utilized to encode each word as a 300-dimensional feature vector. Then, we could represent each image with a 4,096-dimensional feature vector, and each document with an $m\times300$ matrix, where m denotes the maximum word count of the document, and zero-padding was exploited to other documents below this limit. Nevertheless, due to the limitations of our computer, the baselines are unable to directly handle the high-dimensional text features on the XMediaNet dataset. Therefore, we adopted [21] to compute the 300-dimensional mean vector of the $m\times300$ feature matrix for the XMediaNet dataset.

TABLE I: The statistics for the five datasets. In the table below, "*/*/*" represents the size of the training/validation/test subsets in the "Instance" column. The abbreviations for each view are denoted as follows: EN for English, FR for French, GE for German, IT for Italian, and SP for Spanish.

Dataset	Instance	View	Dimensionality
Reuters	10,000/4,000/4,758	EN	$21,531\times1$
		FR	$24,892\times1$
		GR	$34,251\times1$
		IT	$15,506\times1$
		SP	$11,547\times1$
nMSAD	4,000/800/4,000	MV1	28×28
		MV2	28×28
		SAD	25×13
Pascal Sentence	800/100/100	Image	$4,096\times1$
		Text	102×300
XMediaNet	32,000/4,000/4,000	Image	$4,096\times1$
		Text	849×300
MS-COCO	82,081/10,000/30,137	Image	$4,096\times1$
		Text	126×300

2) Evaluation Metric and Baselines: To comprehensively evaluate the effectiveness of our DMLPA, we conducted cross-view retrieval tasks on the Reuters, Pascal Sentence, XMediaNet, and MS-COCO datasets, as well as cross-view recognition tasks on the nMSAD dataset. In cross-view retrieval, we used data from one view as a database and the data from the other views as queries, resulting in $v\times(v-1)$ evaluations based on Mean Average Precision (MAP), as described in [11]. For cross-view recognition, we used data of one view as a gallery and data of other views as probes, leading to $v\times(v-1)$ evaluations based on rank-1 recognition rate, as described in [17]. It is worth noting that we calculated MAP scores on all the returned results in our experiments, following the method in [59].

We compare DMLPA with several related methods, including MCCA [14], GMMFA [11], MvDA [17], MvDA-

VC [6], GSS-SL [21], JRL [20], JGRHML [19], ACMR [40], MCSM [42], CCL [25], CBT [59], SMLN [9], CM-GAN [13], CMCL [2], FedCMR [1], MARS [10], JFSE [60], and SCL [61] . For most linear methods, the reduced dimensionality is determined by the best performance achieved on the validation set traversing the range [1:250] for each dataset, while the other necessary parameters are provided by the original authors. Notably, the results of MCSM, CCL, CBT, CM-GAN, and JFSE come from the original papers. Moreover, for MCSM, CCL, CBT, and CM-GAN, the image and text inputs are extracted by the VGGNet and Sentence CNN fine-tuned on the corresponding training data.

3) Implementation Details: In our experiments, CNN is utilized to handle matrix inputs such as text, audio, and images. Following [13], the text CNN architecture employed in our DMLPA is with the same configuration as [62]. The ReLU activation function [63] is used in all layers, except for the last layer, which adopts the linear activation function. In the experiments, the Euclidean distance is exploited to measure the similarity, *i.e.* $d(\mathbf{y}_i, \mathbf{y}_j) = \|\mathbf{y}_i - \mathbf{y}_j\|$. Learning rate β, maximum epoch, α_1, α_2 and τ are respectively set as 0.001, 100, 0.5, 0.05 and 0.5 in all the experiments on all datasets.

B. Comparisons with State-of-the-art Methods

In this section, we present a comparative analysis of our DMLPA against 18 state-of-the-art approaches on five multi-view datasets. Firstly, we present the MAP scores of 20 retrieval tasks along with their averages on the Reuters dataset, as shown in Table II. Additionally, we showcase the recognition accuracy of 6 recognition tasks and their average accuracy on the nMSAD dataset in Table III. The proposed approach demonstrates significant improvements over the state-of-the-art methods, with an average MAP score increase from 0.814 to 0.821 on the Reuters dataset and an average accuracy increase from 0.899 to 0.976 on the nMSAD dataset. However, it is noteworthy that most of the cross-view methods are not equipped to handle multi-view data that involves more than two views. Therefore, in our comparative analysis, we have only compared the proposed approach with the seven multi-view methods. It is important to highlight that GMMFA, MvDA, MvDA-VC, SMLN, CMCL, and MARS incorporate label information, leading to better performance compared to the unsupervised method MCCA. Furthermore, GMMFA, which takes into account local graphic discriminant information, achieves the best results amongst the shallow methods on the Reuters dataset. However, due to the inability of shallow methods to extract high nonlinear features from the multi-view data, all deep methods outperform the shallow methods remarkably. In addition, our DMLPA outperforms all the methods by considering the intrinsic structure of the input data and the high nonlinear characteristics it exhibits.

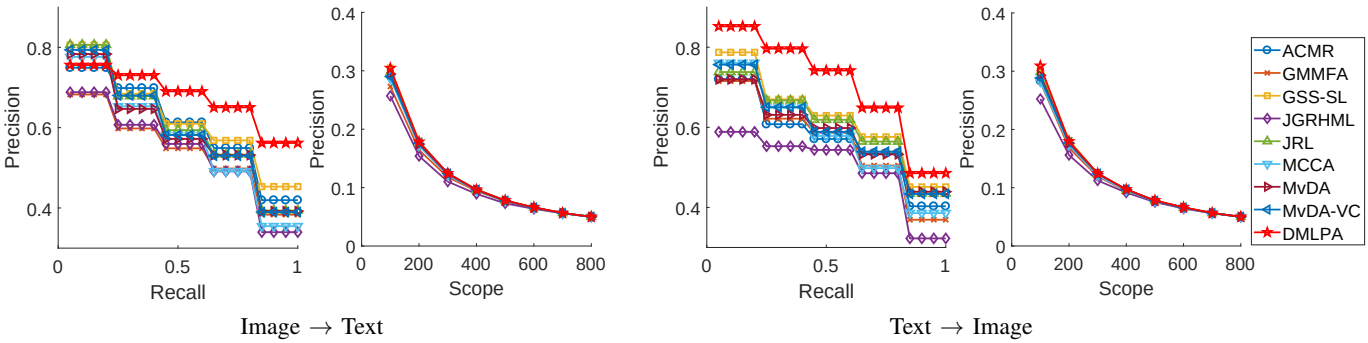
Moreover, Tables IV to VI display the comparative results in terms of the MAP scores on two cross-view retrieval tasks: image query text (Image $\rightarrow$ Text) and text query image (Text $\rightarrow$ Image) on the Pascal Sentence, XMediaNet, and MS-COCO datasets. The experimental results indicate that our DMLPA achieves superior retrieval performance in comparison to 18 state-of-the-art methods on all three datasets. For instance, on the Pascal Sentence dataset, our approach improves the average MAP score from 0.679 to 0.726 in contrast to the best results of its counterpart (i.e., FedCMR). Similar findings can be observed on the XMediaNet dataset from Table V,

TABLE II: Comparative results (MAP@ALL) for cross-language retrieval on the Reuters dataset.

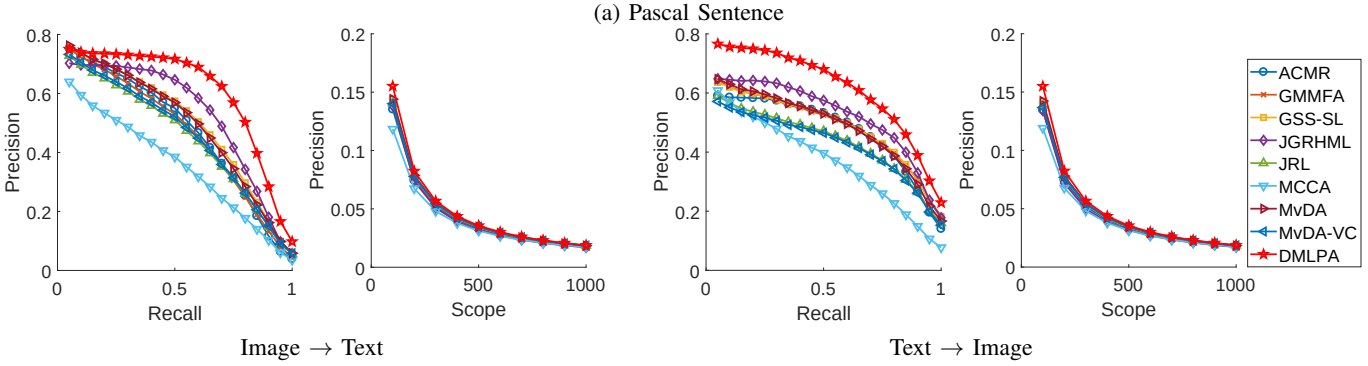
Methods	Query	EN	FR	GE	IT	SP	Avg.
MCCA [14]	EN	-	0.424	0.422	0.423	0.421	0.422
	FR	0.424	-	0.420	0.420	0.418	0.420
	GE	0.422	0.419	-	0.418	0.416	0.419
	IT	0.422	0.419	0.418	-	0.416	0.419
	SP	0.421	0.418	0.417	0.417	-	0.418
	Avg.	0.422	0.420	0.419	0.419	0.418	0.420
GMMFA [11]	EN	-	0.722	0.716	0.718	0.719	0.719
	FR	0.719	-	0.696	0.699	0.700	0.704
	GE	0.714	0.698	-	0.694	0.695	0.700
	IT	0.716	0.700	0.693	-	0.697	0.702
	SP	0.714	0.699	0.692	0.695	-	0.700
	Avg.	0.716	0.705	0.699	0.701	0.703	0.705
MvDA [17]	EN	-	0.637	0.604	0.603	0.630	0.618
	FR	0.648	-	0.658	0.662	0.676	0.661
	GE	0.610	0.655	-	0.604	0.657	0.632
	IT	0.607	0.653	0.601	-	0.647	0.627
	SP	0.637	0.672	0.655	0.651	-	0.654
	Avg.	0.625	0.654	0.630	0.630	0.652	0.638
MvDA-VC [6]	EN	-	0.637	0.566	0.641	0.588	0.608
	FR	0.637	-	0.642	0.585	0.651	0.629
	GE	0.578	0.658	-	0.678	0.653	0.642
	IT	0.625	0.595	0.669	-	0.619	0.627
	SP	0.589	0.656	0.650	0.618	-	0.628
	Avg.	0.607	0.637	0.632	0.631	0.628	0.627
SMLN [9]	EN	-	0.794	0.782	0.787	0.784	0.787
	FR	0.784	-	0.763	0.773	0.771	0.773
	GE	0.755	0.747	-	0.743	0.737	0.745
	IT	0.777	0.772	0.757	-	0.765	0.768
	SP	0.771	0.766	0.753	0.761	-	0.763
	Avg.	0.772	0.770	0.764	0.766	0.764	0.767
CMCL [2]	EN	-	0.821	0.817	0.819	0.819	0.819
	FR	0.822	-	0.813	0.815	0.816	0.817
	GE	0.815	0.810	-	0.808	0.808	0.810
	IT	0.816	0.812	0.808	-	0.810	0.812
	SP	0.817	0.812	0.808	0.810	-	0.812
	Avg.	0.818	0.814	0.812	0.813	0.813	0.814
MARS [10]	EN	-	0.807	0.812	0.808	0.813	0.810
	FR	0.811	-	0.806	0.803	0.804	0.806
	GE	0.807	0.801	-	0.797	0.798	0.801
	IT	0.801	0.792	0.792	-	0.792	0.794
	SP	0.796	0.784	0.786	0.781	-	0.786
	Avg.	0.804	0.796	0.799	0.797	0.802	0.800
DMLPA	EN	-	**0.831**	**0.826**	**0.827**	**0.825**	**0.827**
	FR	**0.833**	-	**0.822**	**0.824**	**0.821**	**0.825**
	GE	**0.829**	**0.822**	-	**0.819**	**0.817**	**0.822**
	IT	**0.829**	**0.823**	**0.818**	-	**0.817**	**0.822**
	SP	**0.827**	**0.821**	**0.816**	**0.817**	-	**0.820**
	Avg.	**0.830**	**0.824**	**0.821**	**0.822**	**0.820**	**0.823**

TABLE III: Comparative results (top-1 accuracy) for cross-view recognition on the nMSAD dataset.

Method	Gallery	SAD		MV1		MV2		Avg.
	Probe	MV1	MV2	SAD	MV2	SAD	MV1	
MCCA [14]		0.569	0.622	0.696	0.531	0.753	0.526	0.616
GMMFA [11]		0.572	0.674	0.711	0.662	0.826	0.623	0.678
MvDA [17]		0.601	0.690	0.788	0.626	0.876	0.604	0.698
MvDA-VC [6]		0.647	0.704	0.820	0.655	0.890	0.666	0.731
SMLN [9]		0.934	0.971	0.924	0.962	0.866	0.736	0.899
CMCL [2]		0.902	0.804	0.981	0.806	0.936	0.851	0.880
MARS [10]		0.915	0.730	0.979	0.567	0.988	0.911	0.848
DMLPA		**0.971**	**0.967**	**0.991**	**0.964**	**0.993**	**0.970**	**0.976**



(a) Pascal Sentence



(b) XMediaNet

Fig. 3: Precision-recall and precision-scope curves for cross-modal retrieval on the Pascal Sentence and XMediaNet datasets.

TABLE IV: Comparative results (MAP@ALL) for cross-view retrieval on the Pascal Sentence dataset.

Method	Image → Text	Text → Image	Avg.
MCCA [14]	0.571	0.574	0.573
GMMFA [11]	0.544	0.565	0.554
MvDA [17]	0.583	0.591	0.587
MvDA-VC [6]	0.568	0.582	0.575
GSS-SL [21]	0.624	0.623	0.623
JRL [20]	0.601	0.605	0.603
JGRHML [19]	0.571	0.562	0.567
ACMR [40]	0.606	0.567	0.587
MCSM [42]	0.598	0.598	0.598
CCL [25]	0.576	0.561	0.569
CBT [59]	0.602	0.583	0.592
CM-GAN [13]	0.603	0.604	0.604
SMLN [9]	0.671	0.644	0.658
CMCL [2]	0.599	0.571	0.585
FedCMR [1]	0.661	0.696	0.679
MARS [10]	0.643	0.621	0.632
JFSE [60]	0.632	0.610	0.621
SCL [61]	0.592	0.611	0.602
DMLPA	**0.723**	**0.729**	**0.726**

TABLE V: Comparative results (MAP@ALL) for cross-view retrieval on the XMediaNet dataset.

Method	Image → Text	Text → Image	Avg.
MCCA [14]	0.361	0.374	0.368
GMMFA [11]	0.472	0.491	0.482
MvDA [17]	0.502	0.491	0.496
MvDA-VC [6]	0.467	0.431	0.449
GSS-SL [21]	0.505	0.493	0.499
JRL [20]	0.460	0.439	0.449
JGRHML [19]	0.535	0.531	0.533
ACMR [40]	0.479	0.519	0.528
MCSM [42]	0.540	0.550	0.545
CCL [25]	0.537	0.528	0.533
CBT [59]	0.577	0.575	0.576
CM-GAN [13]	0.567	0.551	0.559
SMLN [9]	0.584	0.614	0.599
CMCL [2]	0.211	0.248	0.230
FedCMR [1]	0.428	0.351	0.390
MARS [10]	0.645	0.624	0.634
JFSE [60]	0.701	0.691	0.696
SCL [61]	0.310	0.269	0.290
DMLPA	**0.712**	**0.702**	**0.707**

where our DMLPA improves the MAP scores from 0.701 to 0.712 on the image query text task and from 0.691 to 0.702 on the text query image task, as compared to the best baseline JFSE. As some single-label methods (e.g., MvDA, MvDA-VC, GMMFA) cannot be applied on the multi-label MS-COCO dataset, we excluded them from Table VI. In Table VI, the experimental results demonstrate that the supervised methods (i.e., GSS-SL, ACMR, MARS, and DMLPA) outperform the unsupervised methods, indicating the significance of label information in the category-based cross-view retrieval. Thanks to the graph structure, our DMLPA achieves the best performance.

In addition to evaluating our method in terms of the mean average precision (MAP) score, we also compared the precision-recall and precision-scope curves on Pascal Sentence and XMediaNet, as shown in Figure 3. Our DMLPA consistently outperforms all other approaches in terms of precision-recall and precision-scope evaluations, demonstrating its effectiveness for cross-view retrieval. Furthermore, the experimental results in Figure 3, Tables IV and V reveal that

TABLE VI: Comparative results (MAP@ALL) for cross-view retrieval on the MS-COCO dataset.

Method	Image → Text	Text → Image	Average
GSS-SL [21]	0.707	0.702	0.705
ACMR [37]	0.692	0.687	0.690
DCCA [[39]]	0.415	0.414	0.415
DCCAE [37]	0.412	0.411	0.411
TBNN [64], [65]	0.617	0.597	0.607
2WayNet [66]	0.499	0.500	0.500
FedCMR [1]	0.673	0.646	0.646
MARS [10]	0.762	0.750	0.756
SCL [61]	0.725	0.731	0.728
DMLPA	**0.820**	**0.827**	**0.823**

graph regularization methods, such as JRL, JGRHML, and GSS-SL, demonstrate outstanding performance compared to other methods. However, these traditional graph regularization methods require a significant amount of memory space (over 100GB) to handle large-scale datasets like XMediaNet. This indicates that it is expensive for these methods to handle such large-scale datasets.

C. Effect of Pairwise Distance Information

In order to study the effect of pairwise distance information in our DMLPA, we developed and assessed one variation of DMLPA: DMLPA using binary similarity denoted as biDMLPA, which the *apriori* similarity graph matrix $\mathbf{V}$ is a binary matrix, in which $V_{ij} = 1$ if the i-th and the j-th samples have the same class and $V_{ij} = 0$ otherwise. Table VII shows the comparison results in terms of MAP in the same experimental environment on the Pascal Sentence dataset. From the results of Table VII, we can see that DMLPA outperforms biDMLPA, indicating that the pairwise distance information is beneficial for extracting powerful representations from the multi-view data.

TABLE VII: Effect comparison of the pairwise distance information in terms of MAP scores on the Pascal Sentence dataset.

Method	Image → Text	Text → Image	Average
biDMLPA	0.691	0.681	0.686
DMLPA	**0.723**	**0.729**	**0.726**

D. Parameter Analysis

1) Distance Metric Analysis: To investigate the influence of the distance metric $d(\cdot, \cdot)$, comparison experiments are conducted by adopting different distance metrics on the Pascal Sentence dataset, *i.e.*, Chebyshev, Braycurtis, Cosine, Correlation, and Euclidean distances. The experimental results demonstrate the effectiveness of the proposed method with distinct metric distances as shown in Table VIII. From the results, one can see that our DMLPA achieves the best retrieval performance with Euclidean distance. Thus, in this work, we adopt the Euclidean distance to compute the similarity.

2) Convergence Analysis: We also investigate the convergence of our DMLPA on the Pascal Sentence dataset. Figure 4 illustrates the losses versus the distinct number of epochs

TABLE VIII: Comparative results (MAP@ALL) for cross-view retrieval in terms of different distance metrics on the Pascal Sentence dataset.

Method	Image → Text	Text → Image	Average
DMLPA (Braycurtis)	0.664	0.641	0.653
DMLPA (Chebyshev)	0.603	0.612	0.608
DMLPA (Correlation)	0.703	0.709	0.706
DMLPA (Cosine)	0.720	0.715	0.717
DMLPA (Euclidean)	**0.723**	**0.729**	**0.726**

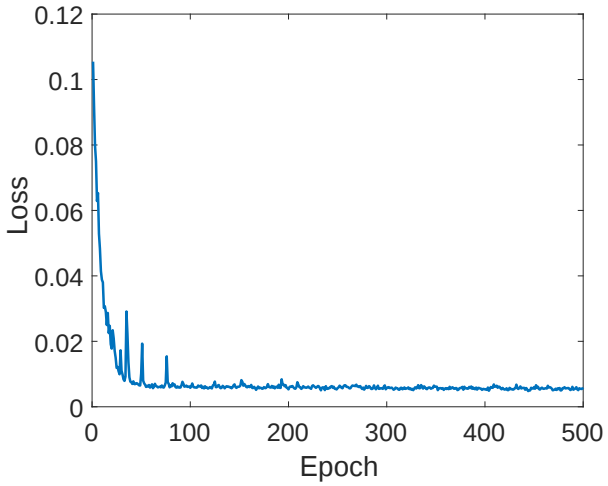


Fig. 4: Convergence analysis. It shows the losses vs. different numbers of epochs on the Pascal Sentence dataset

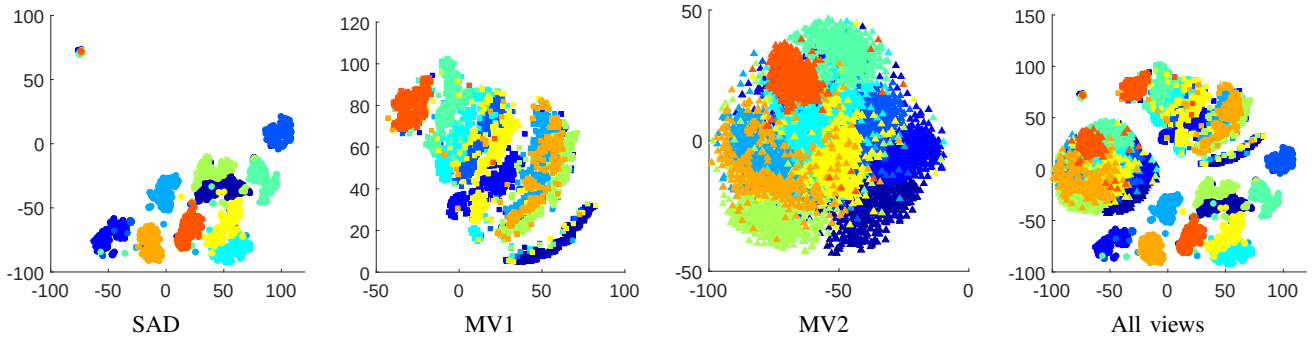
on the Pascal Sentence dataset. The figure reveals that our DMLPA converges within $100 \sim 200$ epochs. As a result, we set the maximal epoch as 200 in the experiments.

E. Time and Memory Cost Analysis

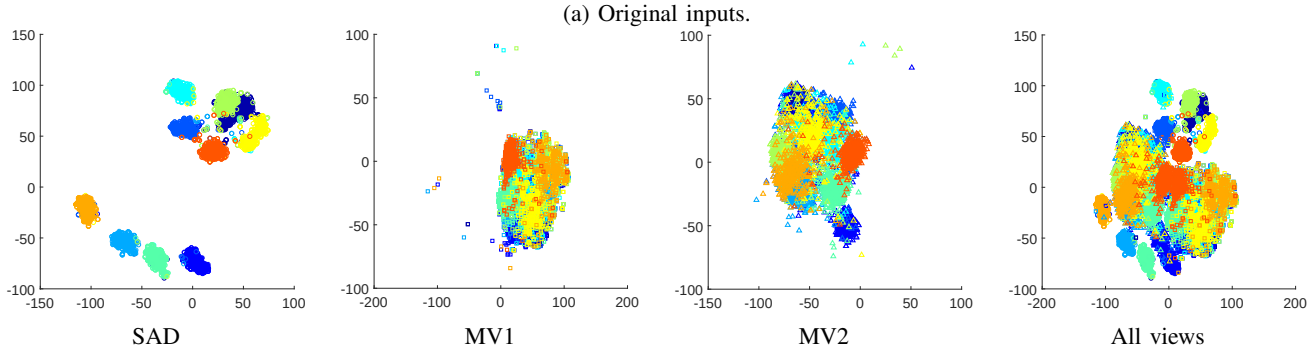
In this section, we compare the computational efficiency of various cross-view graph approaches on the noisy MNIST dataset. We calculate all training time and memory usage on a PC equipped with a 3.50GHz i7-7800X CPU, 128GB RAM, and 64GB SWAP. To ensure a fair comparison, we run all methods on the CPU and set the maximum epochs of DMLPA and JRL to 5, while the maximal epoch of GSS-SL is set to its default value of 2. We present the time and memory cost for each method in Table IX, and observe that the proposed method achieves better efficiency than the others, with remarkably lower time cost and memory usage. Specifically, Table IX reveals that regularization methods require more memory, which limits their ability to handle large-scale multi-view data. Our DMLPA can handle large-scale multi-view data efficiently.

F. Visualisation of the Learned Representation

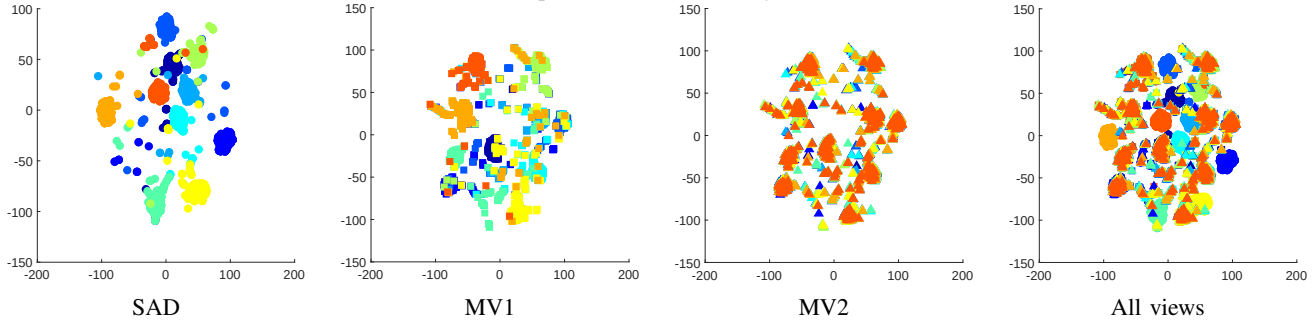
To visually assess the effectiveness of our DMLPA, we employ the t-SNE approach to embed the samples of the three views from the nMSAD dataset into a two-dimensional visualization plane in the common representation space. Figure 5a illustrates the original data distribution, revealing that distinct



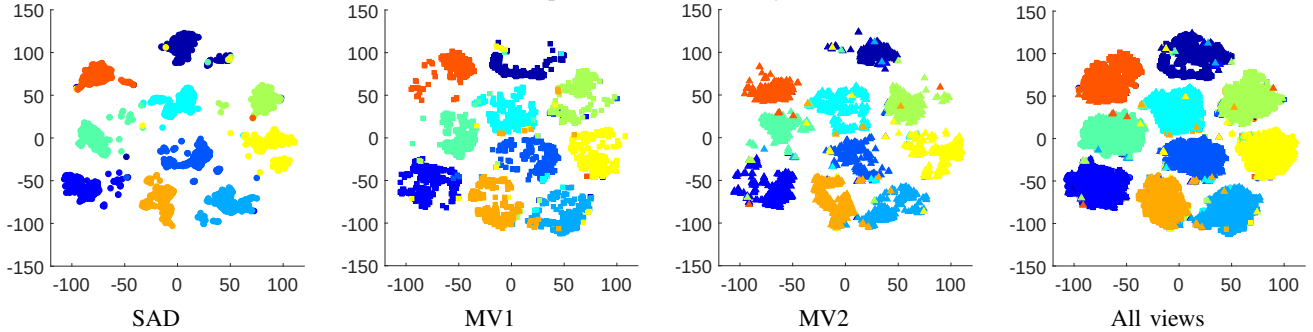
(a) Original inputs.



(b) Common representations learned by GMMFA [11].



(c) Common representations learned by MARS [10].



(d) Common representations learned by our DMLPA.

Fig. 5: This figure shows a visualization of the test data from the nMSAD dataset using the *t*-SNE method. The markers in distinct shapes represent different views, while the distinct colors represent diverse classes. The first, second, third, and fourth columns represent SAD, MV1, MV2, and all views, respectively.

TABLE IX: Efficiency comparison in terms of training time and memory cost on the noisy MNIST dataset.

Data Size	Method	Training Time	Memory (RAM + SWAP)
30K pairs	JGRHML [19]	4651.53s	More than 128GB
	JRL [20]	558.29s	About 52GB
	GSS-SL [21]	2033.72s	About 64GB
	DMLPA	127.95s	About 0.77GB
50K pairs	JGRHML [19]	-	Out of memory
	JRL [20]	1464.75s	More than 128GB
	GSS-SL [21]	9134.52s	More than 128GB
	DMLPA	210.68s	About 0.90GB

views occupy distinct spaces and samples from different classes are not well-separated. In contrast, Figures 5b to 5d depict the data distribution in the learned common space, where it is evident that our DMLPA has effectively projected the distinct views into a shared space, resulting in well-separated samples from diverse classes. Additionally, the learned representations of distinct views exhibit similar distributions and can overlap with each other, indicating that our DMLPA can eliminate view discrepancies while maintaining discrimination in the common space. These observations are consistent with the MAP scores for cross-view recognition and retrieval tasks, where our method outperforms other approaches.

V. Conclusion

In this paper, we present a novel approach for deep multi-view representation learning, termed DMLPA. Our DMLPA aims to approximate an *apriori* similarity graph, rather than relying on a graph regularizer to preserve the graphic information from the input spaces into the common space. The proposed method first constructs a similarity graph from the multi-view inputs. Then, by approximating the constructed similarity graph, the networks can learn to bridge the heterogeneous gap and preserve the discrimination and intrinsic structure of distinct views in the shared space. Furthermore, our DMLPA can be trained in a batch-wise manner, making it more tractable for large-scale multi-view datasets compared to existing cross-view graph regularization methods. Experimental results demonstrate the superiority of our DMLPA over 18 cross-view learning approaches on five datasets *w.r.t.* the cross-view recognition and retrieval tasks.

References

[1] L. Zong, Q. Xie, J. Zhou, P. Wu, X. Zhang, and B. Xu, "Fedcmr: Federated cross-modal retrieval," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1672–1676.

[2] L. Jing, E. Vahdani, J. Tan, and Y. Tian, "Cross-modal center loss for 3d cross-modal retrieval," in *Proceedings of the IEEE/CVF Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 3142–3151.

[3] S. Qian, D. Xue, Q. Fang, and C. Xu, "Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4794–4811, 2023.

[4] P. Hu, Z. Huang, D. Peng, X. Wang, and X. Peng, "Cross-modal retrieval with partially mismatched pairs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9595–9610, 2023.

[5] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo, "A cross-media model for automatic image annotation," in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 73.

[6] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 188–194, 2015.

[7] X. Lu, L. Zhu, Z. Cheng, L. Nie, and H. Zhang, "Online multi-modal hashing with dynamic query-adaption," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 715–724.

[8] L. Zhen, P. Hu, X. Peng, R. S. M. Goh, and J. T. Zhou, "Deep multimodal transfer learning for cross-modal retrieval," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 798–810, 2022.

[9] P. Hu, H. Zhu, X. Peng, and J. Lin, "Semi-supervised multi-modal learning with balanced spectral decomposition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 99–106.

[10] Y. Wang and Y. Peng, "Mars: Learning modality-agnostic representation for scalable cross-media retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4765–4777, 2022.

[11] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2160–2167.

[12] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1234–1244, 2017.

[13] Y. Peng, J. Qi, and Y. Yuan, "CM-GANs: Cross-modal generative adversarial networks for common representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2018.

[14] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, 2010, pp. 1–4.

[15] A. Sharma and D. W. Jacobs, "Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 593–600.

[16] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf, "Randomized nonlinear component analysis," in *International Conference on Machine Learning*, 2014, pp. 1359–1367.

[17] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *European Conference on Computer Vision*, 2012, pp. 808–821.

[18] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *Proceedings of the IEEE Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016, pp. 4847–4855.

[19] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 27, no. 1, 2013, pp. 1198–1204.

[20] ——, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 965–978, 2014.

[21] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 128–141, 2017.

[22] J. Wen, Z. Zhang, Z. Zhang, L. Fei, and M. Wang, "Generalized incomplete multiview clustering with flexible locality structure diffusion," *IEEE Transactions on Cybernetics*, vol. 51, no. 1, pp. 101–114, 2021.

[23] C. Zheng, L. Zhu, Z. Zhang, J. Li, and X. Yu, "Efficient semi-supervised multimodal hashing with importance differentiation regression," *IEEE Transactions on Image Processing*, vol. 31, pp. 5881–5892, 2022.

[24] X. Yang, C. Deng, Z. Dang, and D. Tao, "Deep multiview collaborative clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 516–526, 2023.

[25] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 405–420, Feb 2018.

[26] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 686–701.

[27] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, 2003, pp. 857–864.

[28] A. Globerson and S. T. Roweis, "Metric learning by collapsing classes," in *Advances in Neural Information Processing Systems*, 2006, pp. 451–458.

[29] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[30] L. Maaten, "Learning a parametric embedding by preserving local structure," in *Artificial Intelligence and Statistics*, 2009, pp. 384–391.

[31] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[32] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Transactions on Image Processing*, vol. 11, no. 3, pp. 293–305, 2002.

[33] S. Akaho, "A kernel method for canonical correlation analysis," in *International Meeting of Psychometric Society*, 2001, pp. 263–269.

[34] W. Wang and K. Livescu, "Large-scale approximate kernel canonical correlation analysis," in *International Conference on Learning Representations (ICLR)*, 2016.

[35] X. Peng, J. Feng, J. Lu, W.-Y. Yau, and Z. Yi, "Cascade subspace clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*. SFO, USA: AAAI, Feb. 2017, pp. 2478–2484.

[36] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang, "Structured autoencoders for subspace clustering," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5076–5086, 2018.

[37] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International Conference on Machine Learning*, 2015, pp. 1083–1092.

[38] P. Hu, H. Zhu, J. Lin, D. Peng, Y.-P. Zhao, and X. Peng, "Unsupervised contrastive cross-modal hashing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3877–3889, 2023.

[39] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[40] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the ACM on Multimedia Conference*. ACM, 2017, pp. 154–162.

[41] P. Hu, X. Peng, H. Zhu, J. Lin, L. Zhen, W. Wang, and D. Peng, "Cross-modal discriminant adversarial network," *Pattern Recognition*, vol. 112, p. 107734, 2021.

[42] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Transactions on Image Processing*, pp. 1–1, 2018.

[43] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[44] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*, 2004, pp. 153–160.

[45] N. Quadrianto and C. H. Lampert, "Learning multi-view neighborhood preserving projections," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 425–432.

[46] D. Zhai, H. Chang, S. Shan, X. Chen, and W. Gao, "Multiview metric learning with global consistency and local smoothness," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, p. 53, 2012.

[47] X. Cai, F. Nie, W. Cai, and H. Huang, "Heterogeneous image features integration via multi-modal semi-supervised learning model," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1737–1744.

[48] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[49] M. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views-an application to multilingual text categorization," in *Advances in Neural Information Processing Systems*, 2009, pp. 28–36.

[50] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[51] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[52] N. Hammami and M. Sellam, "Tree distribution classifier for automatic spoken arabic digit recognition," in *International Conference for Internet Technology and Secured Transactions (ICITST)*. IEEE, 2009, pp. 1–4.

[53] N. Hammami and M. Bedda, "Improved tree model for arabic speech recognition," in *International Conference on Computer Science and Information Technology*, vol. 5. IEEE, 2010, pp. 521–526.

[54] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 139–147.

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[56] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 7–16.

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, 2015.

[58] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[59] J. Qi and Y. Peng, "Cross-modal bidirectional translation via reinforcement learning," in *IJCAI*, 2018, pp. 2630–2636.

[60] X. Xu, K. Lin, Y. Yang, A. Hanjalic, and H. T. Shen, "Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3030–3047, 2022.

[61] Y. Liu, J. Wu, L. Qu, T. Gan, J. Yin, and L. Nie, "Self-supervised correlation learning for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 25, pp. 2851–2863, 2023.

[62] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

[63] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, ser. ICML'10. USA: Omnipress, 2010, pp. 807–814.

[64] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016, pp. 5005–5013.

[65] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 394–407, 2018.

[66] A. Eisenschtat and L. Wolf, "Linking image and text with 2-way nets," in *Proceedings of the IEEE Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 4601–4611.