# Dynamic guided metric representation learning for multi-view clustering

**Tingyi Zheng** [Corresp., 1, 2] , **Yilin Zhang** [3] , **Yuhang Wang** [4]

[1] College of Information and Computer, Taiyuan University of Technology, Taiyuan, Shanxi, China

[2] Department of Electrical and control Engineering, Shanxi Institute of Energy, Jinzhong, Shanxi, China

[3] Software College, Taiyuan University of Technology, Taiyuan, Shanxi, China

[4] College of Data Science, Taiyuan University of Technology, Taiyuan, Shanxi, China

Corresponding Author: Tingyi Zheng
Email address: tyut66666@163.com

Multi-view clustering (MVC) is a mainstream task that aims to divide objects into meaningful groups from different perspectives. The quality of data representation is the key issue in MVC. A comprehensive meaningful data representation should be with the discriminant characteristics in a single view and the correlation of multiple views. Considering this, a novel framework called Dynamic Guided Metric Representation Learning for Multi-View Clustering (DGMRL-MVC) is proposed in this paper, which can cluster multi-view data in a learned latent discriminated embedding space. Specifically, in the framework, the data representation can be enhanced by multi-steps. Firstly, the class separability is enforced with Fisher Discriminant Analysis (FDA) within each single view, while the consistence among different views is enhanced based on Hilbert-Schmidt independence criteria (HSIC). Then the 1st enhanced representation is obtained. In the second step, a dynamic routing mechanism is introduced, in which the location or direction information is added to fulfil the expression. After that, a generalized canonical correlation analysis (GCCA) model is used to get the final ultimate common discriminated representation. The learned fusion representation can substantially improve multi-view clustering performance. Experiments validated the effectiveness of the proposed method for clustering tasks.

1 **Dynamic Guided Metric Representation Learning for Multi-**
2 **view Clustering**

3 Tingyi Zheng[1,2], Yilin Zhang[3], Yuhang Wang[4]

4 [1] College of Information and Computer, Taiyuan University of Technology, Taiyuan/Shanxi, China

5 [2] Department of Electrical and control Engineering, Shanxi Institute of Energy, Jinzhong/Shanxi, China

6 [3] Software college, Taiyuan University of Technology, Taiyuan/Shanxi, China

7 [4] College of Data Science, Taiyuan University of Technology, Taiyuan/Shanxi, China

9 Corresponding Author:

10 Tingyi Zheng[1,2]

11 No.79 West Street Yingze, Taiyuan, Shanxi, 030024, China

12 Email address: tyut66666@163.com

## Abstract

13            Multi-view clustering (MVC) is a mainstream task that aims to divide objects into meaningful
14   groups from different perspectives. The quality of data representation is the key issue in MVC. A
15   comprehensive meaningful data representation should be with the discriminant characteristics in a
16   single view and the correlation of multiple views. Considering this, a novel framework called
17   Dynamic Guided Metric Representation Learning for Multi-View Clustering (DGMRL-MVC) is
18   proposed in this paper, which can cluster multi-view data in a learned latent discriminated
19   embedding space. Specifically, in the framework, the data representation can be enhanced by
20   multi-steps. Firstly, the class separability is enforced with Fisher Discriminant Analysis (FDA)
21   within each single view, while the consistence among different views is enhanced based on
22   Hilbert-Schmidt independence criteria (HSIC). Then the 1st enhanced representation is obtained.
23   In the second step, a dynamic routing mechanism is introduced, in which the location or direction
24   information is added to fulfil the expression. After that, a generalized canonical correlation
25   analysis (GCCA) model is used to get the final ultimate common discriminated representation.
26   The learned fusion representation can substantially improve multi-view clustering performance.
27   Experiments validated the effectiveness of the proposed method for clustering tasks.

## Introduction

Multi-view data are extremely common in many applications, each individual view and the correlation of multiple views have their specific property for a particular knowledge discovery task. Multi-view data often contain diversity as well as consistent information that should be exploited and fused. Therefore, considering the uniqueness, complementary, and correlation of each view, it is essential to study how to fuse multi-view data efficiently. Recently, increasing research efforts have been made in multi-view learning, where multi-view clustering (MVC) forms a mainstream task that aims to divide subjects into meaningful groups from different perspectives by learning the multi-view information (e.g. Yang & Wang,2018). However, a better clustering representation can be learned by simultaneously analyzing the discriminant characteristics of a single view and the correlations among multiple views. Therefore, a priority for most MVC methods is to find a feasible and direct way to explore the underlying data cluster latent fusion representation by multiple views to obtain the final ideal clustering results.

Many research work about MVC have been investigated in the last decades, in which most of them proposed to effectively consider rich information from multiple views (e.g. Wang, Yang &Liu,2020). For instance, co-training (e.g. Xia, Yang & Yu, 2020) and co-regularized (e.g. Kang et al.,2020) spectral clustering have been proposed to minimize the disagreement between each pair of views (e.g. Chen et al.,2020 ). However, clustering performance is easily affected by the poor quality of original views in this kind of method. Multi-view subspace clustering uses the unified shared feature representation of each view to obtain consistent clustering results from multiple views (e.g. Huang et al.,2020; Zhao, Ding & Fu, 2017). Typical models also include subspace learning-based and non-negative matrix factorization (NMF)-based models. In the past decade, numerous machine learning technologies have been investigated to determine the scope of combining multiple views. However, these methods lack the ability to mine the latent unified representation and to learn the non-linear correlation of views. To address the second limitation, many researchers proposed multi-kernel and CCA-based clustering methods. The multi-kernel

55  method uses predefined kernels for each view and combines these kernels in linear or non-linear
56  mode. Nevertheless, the complex relationships make it difficult to represent data fusion. To solve
57  this problem, canonical correlation analysis (CCA) (e.g. Haldorai &Ramu,2020; Hotelling,1936)
58  and kernel CCA are commonly used in multi-view clustering. Rasiwasia et al. proposed mean
59  CCA and cluster CCA. Blaschko (e.g. Rasiwasia et al.,2008) projected the data across different
60  views by KCCA and used k-means to cluster projected samples. We can find that the
61  computational cost of the CCA and KCCA models is high (e.g. Wang, Li &Huang, 2017). Besides,
62  existing CCA-based clustering methods are only focus on mining the linear correlations. With the
63  wide application of deep learning, the deep network has been applied in CCA model. Such as,
64  Deep CCA (DCCA) (e.g. Andrew, Arora & Bilmes, 2013), Deep Generalized CCA (DGCCA)
65  (e.g. Benton et al., 2017). However, DCCA model can only learn two views. Although DGCCA
66  overcame the limitation of DCCA, it ignores mining the specific characteristics of multiple views.
67  Subsequently, the newest multi-view representation learning model (MRL) (e.g. Zheng et al.,
68  2020) based on DGCCA is proposed. This model is focus on mining the specific characteristics of
69  inner-view and then learning the fusion representation based on the maximum correlation of
70  multiple views. Recently, multi-view representation based on clustering has been viewed as the
71  problem of learning a meaningful representation of data. Therefore, how to design model for multi-
72  view clustering is an intriguing direction.

73      In summary, an effective mechanism is to learn a comprehensive meaningful representation
74  with the discriminant characteristics in a single view and the correlation of multiple views. DCCA-
75  based multi-view clustering methods are rarely used, but they have room for improvement with
76  deep networks to mine nonlinear correlation and high-level fusion representations. With the aim
77  of addressing the limitation of DCCA-based MVC methods, this paper proposes a unique novel
78  framework called Dynamic Guided Metric Representation Learning for Multi-View Clustering
79  (DGMRL-MVC) that has not been investigated already in previous works on this topic.

80      The main contributions of this work can be summarized as follows:
81      ● A multi-step enhanced representation is proposed for multi-view clustering, which consists
82  of inter-intra learning, deep learning and latent space mapping. The proposed model can jointly
83  learn a latent discriminated embedding.
84      ● On the basis of learning intra-view class separability and inter-view consistency by Fisher
85  Discriminant Analysis -with Hilbert-Schmidt Independence Criteria (FDA-HSIC) metric learning,
86  a dynamic guided deep learning method is used that introduces location or direction information
87  to enhance the single view representation.
88      ● An ultimate common representation is obtained by generalized canonical correlation
89  analysis (GCCA) model for multiple views.
90      ● Experiments on four real-world multi-view datasets have validated the effectiveness of the
91  proposed method for clustering tasks.
92      The rest of this paper is organized as follows: Section 2 presents a review of related work.
93  The next section introduces the proposed DGMRL-MVC model. Section 4 presents the datasets,
94  experimental settings, and experimental results. Finally, Section 5 draws conclusions.

# Related work

Related work on multi-view clustering methods can be divided into two categories: the common matrix framework (spectral clustering, subspace clustering, and non-negative matrix factorization clustering) and view fusion methods (multi-kernel clustering and DCCA-based methods). This section will review multi-view clustering methods from these two technological categories.

## 2.1 Common matrix framework

Methods of this type have the commonality that they share the similar structure to combine multiple views.

### 2.1.1    Multi-view spectral clustering

Multi-view spectral clustering assumes that all the views share the same or similar eigenvector matrices, in which co-training spectral clustering and co-regularized spectral clustering are the two representative methods.

1) Co-training spectral clustering. These algorithms are investigated under the assumption of consensus among multiple views and trained alternately to maximize the consistency of the two distinct views. Three main assumptions are made: (1) each view is sufficient for the learning task; (2) the views are conditionally independent given the class labels; and (3) the objective functions export the same predictions for co-occurring features with high probability in both views. Overall, most co-training methods are semi-supervised learning.

2) Co-regularized spectral clustering. Zhu (e.g. Zhu, Zhang &He, 2019) proposed co-regularized approach. The main idea of this kind of method is to minimize the distinction between the predictor functions of two views acting as one part of an objective function. They used graphic Laplacian eigenvectors to play a role much like that of predictor functions in a semi-supervised learning scenario. Inspired by the previous work (e.g. Zhu, Zhang &He, 2019; Chen et al., 2020), automatically learned the weights of different views from data.

### 2.1.2 Multi-view subspace clustering

In practice, multi-view data can be sampled from multiple subspaces. A subspace learning model can learn a new and unified representation or a latent space for multi-view data. The unified representation or latent space can then be directly used for the clustering task. In addition, a subspace clustering model can deal with high-dimensional data. The approach is to find the underlying subspaces and then to cluster. Wang (e.g. Wang et al.,2015) proposed a model to measure correlation consensus in multiple views. Unlike Wang, Zhao (e.g. Zhao et al.,2017) used a deep semi-nonnegative matrix factorization to perform clustering. This kind of method focus on mining the inherent structure from multiple subspace of views and the clustering performance is heavily dependent on the affinity matrix. Therefore, some works used deep networks to learn the inter-view specific features based on classical subspace clustering methods. The newest multi-view representation learning model (MRL) (e.g. Zheng et al., 2020) is focus on mining the specific characteristics of inner-view and then learning the fusion representation based on the maximum correlation of multiple views. In all, how to mining the hidden difference of views is a research topic in recent years.

### 2.1.3 Multi-view non-negative matrix factorization clustering

Given a non-negative matrix $X \in \mathbb{R}^{d \times n}$, non-negative matrix factorization (NMF) seeks two non-negative matrices $W \in \mathbb{R}^{d \times p}$ and $V \in \mathbb{R}^{p \times n}$, whose product the original matrix X:

$$X \approx WV^{T},$$

where W is the basis matrix and V the indicator matrix.

140      Due to its non-negativity constraints, NMF has emerged as a latent feature learning method.
141 To combine multi-view information in an NMF framework, many variants of NMF have also been
142 proposed. Kalayeh (e.g. Kalayeh, Idrees & Shah,2014) present a weighted extension of Multi-
143 view NMF to address the aforementioned drawbacks. Liu (e.g. Liu et al.,2020) used the cross
144 entropy loss function to constrain the objective function better.

**2.2 View fusion methods**

146      View fusion methods are also commonly used to learn multi-view information. However, their
147 learning targets directly combine the views for clustering by different modes. Hence, they focus on
148 how to combine multiple views.

**2.2.1   Multi-view multi-kernel clustering**

150      Multi-kernel learning was originally used to combine views by means of a kernel and was
151 widely used to deal with multi-view data. The common approach is to define a kernel for each view
152 and then combine these kernels in a convex combination (e.g. Sellami &Alaya, 2021; Lu et
153 al.,2020;Nithya et al.,2020). Therefore, one main problem of this method is to choose kernel
154 functions and optimal. A $k$-means analysis performed on kernel space that simultaneously found
155 the best cluster labels, the cluster kernels, and the optimal combination of multiple kernels (e.g. Cui
156 et al.,2007). Liu (e.g. Liu et al.,2013) extended $k$-means clustering into Hilbert space. Liu try to
157 denote the data matrices as kernel matrices and then combined for fusion. In addition, consider the
158 differences among views, methods with weighted combinations of kernels have also been studied.
159      However, multi-view data are incomplete. To address this issue, researchers integrated kernel
160 imputation and clustering into a unified learning procedure (e.g. Monney et al.,2020). Besides, other
161 methods use a direct combination of features to perform multi-view clustering, like those in (e.g.
162 Xu et al.,2016; Nie et al.,2017).

**2.2.2 Multi-view CCA-based methods clustering**

164      For multi-view data clustering, it is reasonable to combine the data directly. However, it is hard
165 to fuse different data types, and high dimensionality and noise are difficult to handle. The CCA
166 method is used to combine the views directly and CCA-based techniques for multi-view clustering
167 fusion. Chaudhuri (e.g. Chaudhuri et al.,2009 ) projected the data into a lower-dimensional space
168 using CCA and cluster. Then, he CCA-based methods are proposed, such as KCCA (e.g. Wang, Li
169 &Huang, 2017), DCCA(e.g. Andrew, Arora & Bilmes, 2013),DGCCA (e.g. Benton et al., 2017)
170 and MRL(e.g. Zheng et al., 2020). Overall, although DCCA-based methods are used now in multi-
171 view representation, this is a direction worthy of further multi-view clustering study using DCCA-
172 based methods. In contrast to these methods, the representation will directly influence effectiveness
173 on the clustering task. Thus, based on existing multi-view clustering learning methods, the intention
174 is to design a novel model to learn a comprehensive meaningful data representation with the
175 discriminant characteristics in a single view and correlation of all views.
176      In short, existing methods still has the following challenges: (1) Common matrix methods
177 focus on learning a shared similar structure, they ignore the hidden difference of views. (2) View
178 fusion methods considering the unique features of each view and the correlation of all views.
179 However, existing methods strategy is two steps. The first step is self-representation of each view
180 independently and the next step is focus on fusion views. The two-step collaboration is not well
181 meet the clustering task.

# The proposed model

**3.1 Motivation**

184      Existing CCA-based multi-view clustering approaches that can deal with multi-view data learn
185 the correlation of two views and perform a clustering task at the same time. Despite appealing

186  performance, they still have some limitations. First, the quality of data representation is the key
187  issue in MVC. And in real-world applications, directly fusing multi-view data is difficult and the
188  effect is not good. Second, the discriminant characteristics in a single view and the correlation of
189  multiple views should all be considered. Considering this, a novel framework called Dynamic
190  Guided Metric Representation Learning for Multi-View Clustering (DGMRL-MVC) is proposed
191  in this paper, which can cluster multi-view data in a learned latent discriminated embedding space.
192  Specifically, in the framework, the data representation can be enhanced by multi-steps. The model
193  combines the effectiveness of FDA-HSIC and the dynamic routing learning of views embedded in
194  the model. Hence, it can mine the class separability of single view, the consistence among different
195  views, discriminant characteristics in a single view and a latent discriminated embedding for
196  clustering task.

197  **3.2 Problem formulation**

198      Given multi-view data observations, let $X_j \in \mathbb{R}^{d_j \times N}$ denote the $j^{th}$ view, where $d_j$ is the
199  feature dimensionality of the $j^{th}$ view and $N$ is the number of samples.

200  **Definition 1.** (Mahalanobis distance). The Mahalanobis distance between two views $x_1$ and $x_2$ is
201  defined as:

202  $$d_M^2(x_1, x_2) = \|x_1 - x_2\|_M^2 = (x_1 - x_2)^T M (x_1 - x_2), \quad (1)$$

203  where, the Mahalanobis matrix M is constrained to be symmetric positive-definite to ensure its
204  validity.

205  **3.3 Framework**

206  **3.3.1 Network Architecture**

207      Fig.1 illustrates the architecture of the proposed DGMRL-MVC model for multi-view
208  clustering. This architecture effectively improves fusion representation and overcomes the
209  limitations of traditional DCCA-based multi-view clustering learning. The model consists of three
210  modules: inter-intra representation based on Fisher discriminant analysis and the Hilbert-Schmidt
211  independence criteria, together called FDA-FISH metric learning; deep representation based on
212  dynamic guided  deep learning; shared representation based on deep generalized canonical
213  correlation analysis. The first module learns intra-view separability and inter-view consistency.
214  Then, as for dynamic guided deep learning, it is used to introduces location or direction information
215  to enhance the single view representation. In the last sub-model, the fusion representation **G** of
216  multiple views for the clustering task is created.

217  Figure 1: **Illustration of the proposed DGMRL-MVC model.**

218  1) Inter-intra representation based on FDA-HSIC

219      The Fisher-HSIC Multi-View Metric Learning (FISH-MML) system is proposed in the work
220  (e.g. Zhang et al.,2018). This method is based on FDA and HSIC, which are simple, yet rather
221  effective. Inspired by Zhang's work, an effort was made to add FDA-HSIC into the proposed
222  framework for multi-view clustering representation learning.

223      **Intra-view separability.** The starting point is the definitions of between-class and total
224  scatter matrices in the $v^{th}$ view:

225  $$S_b^{(v)} = \frac{1}{n}\sum_{j=1}^{m} n_j\left(\mu_j^{(v)} - \mu^{(v)}\right)\left(\mu_j^{(v)} - \mu^{(v)}\right)^T, (2)$$

226  $$S_t^{(v)} = \frac{1}{n}\sum_{i=1}^{n} (z_i^{(v)} - \mu^{(v)})(z_i^{(v)} - \mu^{(v)})^T, (3)$$

227  $$\mu_j^{(v)} = \frac{1}{n_j}\sum_{i=1}^{n_j} z_i^{j(v)}, \mu^{(v)} = \frac{1}{n}\sum_{i=1}^{n_j} z_i^{(v)}, (4)$$

228 where $b$ is the abbreviation of between, $t$ is the abbreviation of total, $m$ is the number of classes,
229 $n_j$ is the number of samples belonging to class $C_j$, $\mu_j^{(v)}$ are the sample means of the $v^{th}$ view for
230 class $C_j$, and $\mu^{(v)}$ are the sample means of the $v^{th}$ view for all the views. $z_i^{j(v)}$ is the projected feature
231 vector corresponding to $x_i^{j(v)}$ and is defined as follows:

232
$$z_i^{j(v)} = P^{(v)}x_i^{j(v)}, (5)$$

233 where the symmetric positive semi-definite matrix $M = \mathbf{P}^{\mathrm{T}}\mathbf{P}$ and $\mathbf{P} \in \mathbb{R}^{k \times d}(\mathrm{k} \geq rank(\mathrm{M}))$ is the
234 new space.

235 The trace operator for $S_b^{(v)}$ and $S_t^{(v)}$ conditioned on $\mathbf{P}^{(v)}$ is then optimized by the following
236 optimization function:

237
$$\max_{\{P^{(v)}\}_{v=1}^{V}} \Sigma_{v=1}^{V} Tr(\mathbf{S}_b^{(v)};\mathbf{P}^{(v)}) - \gamma Tr(\mathbf{S}_t^{(v)};\mathbf{P}^{(v)}), (6)$$

238 where $\gamma$ is a tunable parameter to balance the two terms involved. Therefore, the above
239 optimization function is focused on seeking metrics that jointly maximize separability.

240 **Inter-view consistency.** The next step is to explore the complementarity information from
241 multi-views using HSIC (e.g. Gretton et al.,2005). Based on HSIC, the function can be defined as:

242
$$\mathrm{HSCI}(\mathbf{Z}^{(v)},\mathbf{Z}^{(w)}) = (n-1)^{-2}\mathrm{tr}(\mathbf{K}_v\mathbf{H}\mathbf{K}_w\mathbf{H}), (7)$$

243
$$\mathbf{K}^{(v)} = \mathbf{Z}^{(v)T}\mathbf{Z}^{(v)} = \mathbf{X}^{(v)T}\mathbf{P}^{(v)T}\mathbf{P}^{(v)}\mathbf{X}^{(v)}, (8)$$

244
$$\mathbf{K}^{(w)} = \mathbf{Z}^{(w)T}\mathbf{Z}^{(w)} = \mathbf{X}^{(w)T}\mathbf{P}^{(w)T}\mathbf{P}^{(w)}\mathbf{X}^{(w)}, (9)$$

245 where $\mathbf{K}^{(v)}$ and $\mathbf{K}^{(w)}$ are the Gram matrices from the two views parameterized by the projections
246 $\mathbf{P}^{(v)}$ and $\mathbf{P}^{(w)}$. In the model, the dependency between $\mathbf{K}^{(v)}$ and $\mathbf{K}^{(w)}$ is enhanced by maximizing the
247 HSIC function, and $\mathbf{H}$ is the Gram matrix that ensures zero mean in the feature space.

248 **Optimization.** The FDA-HSIC optimization function can be described as:

249
$$\max_{\{P^{(v)}\}_{v=1}^{V}} \sum_{v=1}^{V} Tr(\mathbf{S}_b^{(v)};\boldsymbol{P}^{(v)}) + \lambda_1 Tr(\mathbf{S}_t^{(v)};\boldsymbol{P}^{(v)}) + \lambda_2\sum_{v \neq w} \mathrm{HSIC}(\boldsymbol{P}^{(v)}\mathbf{X}^{(v)},\boldsymbol{P}^{(w)}\mathbf{X}^{(w)})$$

250
$$= \max_{\{P^{(v)}\}_{v=1}^{V}} \Sigma_{v=1}^{V} Tr(\boldsymbol{P}^{(v)}(\mathbf{A} + \lambda_1\mathbf{B} + \lambda_2\mathbf{C})\boldsymbol{P}^{(v)\mathrm{T}}$$

251
$$= \max_{\{P^{(v)}\}_{v=1}^{V}} \sum_{v=1}^{V} tr(\mathbf{P}^{(v)}\mathbf{D}\mathbf{P}^{(v)T})$$

252
$$s.t.\ \mathbf{P}^{(v)}\mathbf{P}^{(v)\mathrm{T}} = \mathbf{I}, \mathrm{v} = 1,2,...,\mathrm{V},\ (10)$$
253
$$\text{where } \lambda_1 > 0 \ and \ \lambda_2 > 0.$$

254 2) Deep representation based on dynamic guided learning
255 The output $\mathbf{P}^{(v)}$ of first submodel is fed into this submodel. $\mathbf{P}^{(v)}$ is the $v^{th}$ view. In this process,
256 the learning of each view is independent. Inspired by Sabour's work (e.g. Sabour,Frosst
257 &Hinton,2017), this submode adopt the dynamic guided mechanism to learn the location or
258 direction information of view. Taking the first view $\mathbf{P}^{(1)}$ as example, we first divide $\mathbf{P}^{(1)}$ into several
259 capsules $u_1,...,u_{k1}$. Each capsule $u_{k1}$ by a weight matrix $W_{k_1 k_2}^{*}$ and get a prediction vector $\hat{o}_{k_2|k_1}$.
260 $\hat{o}_{k_2|k_1}$ can be called deep learning mapping. Next, to represent the probability for the output vector,
261 a nonlinear function squashing is applied to $s_{k_2}$. And the new capsule $s_{k_2}$ is weighted sum over $u_{k_1}$
262 with the coupling coefficient $c_{k_1 k_2}$.

263
$$\hat{o}_{k_2|k_1} = W_{k_1 k_2}^{*}u_{k_1}, (11)$$

264
$$s_{k_2} = \Sigma_{k1}\, c_{k_1 k_2}\, \hat{o}_{k_2|k_1}, \quad (12)$$

265
$$c_{k_1 k_2} = \text{softmax}\big(p_{k_1 k_2}\big) = \frac{\exp\,(p^*_{k_1 k_2})}{\Sigma_k \exp\,(p^*_{k_1 k})}, \quad (13)$$

266  where the initial logits $p^*_{k_1 k_2}$ is the log prior probabilities that capsule $k_2$ to $k_1$.

267  Finally, the output of the first view is $\mathbf{P}^{*\,(1)}$, it is denoted as: $\mathbf{P}^{*\,(1)} = [\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k_2}}]$, where $\mathbf{v_{k_2}}$

268  $= [\mathbf{v_1}, \ldots, \mathbf{v_j}]$. The discriminant characteristics of single view is learned:

269
$$\mathbf{v_{k_2}} = \text{squas}h\big(s_{k_2}\big) = \frac{\|s_{k_2}\|^2}{1 + \|s_{k_2}\|^2}\frac{s_{k_2}}{\|s_{k_2}\|}, \quad (14)$$

270  3) Shared representation based on deep generalized canonical correlation analysis
271  This sub-model learns the ultimate common fusion representation $\mathbf{G}$ from all views. In this
272  process, minimized loss function as follows:

273
$$\underset{U_v \in \mathbb{R}^{d_v \times r},\, G \in \mathbb{R}^{r \times N}}{\text{minimize}} \sum_{v=1}^{V} \big\| \mathbf{G} - \mathbf{U_v^T P}^{*\,(v)} \big\|_F^2,$$

274
$$\text{subject to } \mathbf{GG^T} = \mathbf{I_r}, \quad (15)$$

275  where $\mathbf{P}^{*\,(v)} \in \mathbb{R}^{d_v \times N_v}$ is latent discriminated embedding of the $v^{th}$ view. Then, all $\mathbf{P}^*_v$ spliced into
276  a matrix $\mathbf{G}$.

277  **3.3.2 Training**
278  In this model, the scaled empirical co-variance matrix $\mathbf{C_{vv}}$ of the $v^{th}$ network is calculated by
279  $\mathbf{C_{vv}} = \mathbf{P}^{*\,(v)}(\mathbf{P}^{*\,(v)})^\mathbf{T} \in \mathbb{R}^{d_v * d_v}$. Where $\mathbf{P}^{*\,(v)} = \mathbf{P}^{*\,(v)\mathbf{T}}\mathbf{C}_{vv}^{-1}\mathbf{P}^{*\,(v)} \in \mathbb{R}^{N*N}$ is the corresponding
280  projection matrix. In optimization process, we consider maximize the sum of correlations between
281  fusion representation $\mathbf{G}$ and single view. The reconstruction error as follows:

282
$$\sum_{v=1}^{V} \big\| \mathbf{G} - \mathbf{U_v^T P}^{*\,(v)} \big\|_F^2 = \sum_{v=1}^{V} \big\| \mathbf{G} - \mathbf{GP}^{*\,(v)\mathbf{T}}\mathbf{C}_{vv}^{-1}\mathbf{P}^{*\,(v)} \big\|_F^2$$

283
$$= r\mathbf{J} - \text{Tr}(\mathbf{GMG}^T), \quad (16)$$

284  where the positive semi-definite matrix of each view is $\mathbf{M} = \Sigma_{v=1}^{V} \mathbf{P}_v$, $r$ is the top rows of $\mathbf{M}$. To

285  learn the correlation of every pair of views, we stack them in a $j * j$ matrix, it is defined as $\mathbf{J}$.

286  The derivative of loss function is :

287
$$\frac{\partial \mathbf{L}}{\partial \mathbf{P}^*_v} = 2\mathbf{U_v}\mathbf{G} - 2\mathbf{U_v}\mathbf{U_v^T}\mathbf{O}^*_v, \quad (17)$$

288  where $\mathbf{L} = \Sigma_{i=1}^{r} \lambda_i(\mathbf{M})$.

# Experiments

**4.1 Datasets**
291  We select four datasets come from real-world to verify model. The views of these datasets
292  include network, text, IDs and image.
293  ●**Football**. This dataset is collected as previously described in Greene D (2013). It is the
294  Twitter active data of 248 English Premier League football players and clubs. This dataset contains
295  nine views and 20 class labels.

296       ●**Wikipedia**. This dataset is collected as previously described in N.Rasiwasia et al. (2020) &
297 J.Costa P et al. (2014).It come from Wikipedia's featured articles. This dataset contains two views
298 and 10 categories.
299       ●**Handwritten**. This dataset is collected through https://archive.ics.uci.edu/ml/datasets/Mult-
300 Iple+Features. It contains six views and 10 categories.
301       ●**Pascal**. This dataset is collected as previously described in C.Rashtchian et al. (2010). It
302 contains six views and 20 categories.
303 **4.2 Experimental settings**
304    In the experiments, we choose 80% of data randomly to train and other 20% to test. In FDA-
305 HSCI, the parameter tuning was as in Zhang's work (e.g. Zhang et al.,2018). The maximum
306 training iterations and learning rate η as in Beton et al. (2017).
307 **4.3 Evaluation metrics**
308    The effect of our model is evaluated by comparing it with baselines in terms of clustering
309 precision, recall, F1, RI, Normalized Mutual Information(NMI) and Silhouette_score. These
310 evaluation measures are defined as follows:

311
$$precision = \frac{TP}{TP + FP}, (18)$$

312
$$recall = \frac{TP}{TP + FN}, (19)$$

313
$$F1 = \frac{2PR}{P + R}, (20)$$

314
$$RI = \frac{TP + TN}{TP + FP + FN + TN}, (21)$$

315 where TP and FP are positive and negative samples predicted as positive classes, TN and FN are
316 positive and negative samples predicted as negative classes.

317
$$NMI(X;Y) = \frac{2I(X;Y)}{H(X) + H(Y)}, (22)$$

318 where $I(X;Y) = \sum_x \sum_y p(x,y) log \frac{p(x,y)}{p(x)p(y)}, H(X) = -\sum_i p(x_i) log p(x_i)$.

319    The Silhouette_score for a sample is:

320
$$Silhouette\_score = \frac{b - a}{\max(a,b)}, (23)$$

321 where $b$ is the distance between a sample and the nearest cluster of which the sample is not a
322 member. Values close to 1 indicate that a sample has been assigned to a better cluster because a
323 different cluster is more similar.
324 **4.4 Baselines**
325    The following methods were compared in the experiments described in this section:
326       ● **DCCA**: The method realizes the deep learning applied in CCA. However, DCCA can only
327 learn nonlinear mapping of two views (e.g. Andrew, Arora & Bilmes, 2013).
328       ●**DGCCA:** Although this model overcame the view number limitation of DCCA, it ignores
329 mining the specific characteristics of multiple views (e.g. Benton et al., 2017).
330       ●**FISH-MML**: A Fish-HSIC multi-view metric learning method, which is simple, yet rather
331 effective. The model can learn intra-view separability and inter-view correlation (e.g. Zheng et al.,
332 2020).
333       ●**Dynamic guided representation-MVC:** To better verify the validity of the model with
334 FDA-FISH, only the dynamic guided representation was added to the proposed model (e.g. Zheng
335 et al., 2020).

Table 1 shows the differences among these methods. In summary, the effectiveness of the proposed DGMRL-MVC model was directly evaluated by comparing it with newest DCCA-based multi-view learning methods. The significance of the dynamic guided representation mechanism in the proposed model was further verified by comparison with FISH-MML. To provide a further comprehensive analysis of the stability and performance of the proposed multi-view representation learning model, the dynamic guided representation-MVC was also chosen as a baseline.

**Table1:**

**Differences between baselines and the proposed method**

**4.5 Results**

The experimental results answered the following questions:

*(RQ1) How good is the performance improvement of the proposed model with dynamic guided representation? (DGMRL-MVC vs. FISH-MML)*

FISH-MML provided the best multi-view metric learning with FDA-FISH. The proposed DGMRL-MVC model was first compared with FISH-MML. The results showed that DGMRL-MVC significantly outperformed FISH-MML on all four datasets for the multi-view clustering task. Table 2 presents the comprehensive comparison results. It is obvious that the proposed model achieved the best performance on nearly all datasets under all metrics. Take the Handwritten dataset with six views as an example, the improvement of DGMRL-MVC over FISH-MML was about 42.2%, 60.4%, 46%, 60.3% and 19.5% in terms of precision, recall, F1, RI, and NMI respectively. This proved that the dynamic guided representation used in the proposed method is effective for multi-view clustering. Our model has been enhanced in two aspects, i.e., inter-intra relations and dynamic routing learning embedded into each view self-representation, and fusion representation with maximum correlation.

**Table 2:**

**Improvement performance of dynamic guided representation on different views. The best results are highlighted in bold.**

*(RQ2) How good is the performance of DGMRL-MVC compared with other DCCA-based methods?*

The latent DCCA-based representation was compared with the proposed algorithm on a clustering task. DCCA and DGCCA were chosen as baselines. In the dynamic guided representation-MVC model, FDS-FISH was not present, and only dynamic guided representation and generalized canonical correlation analysis were added. As shown in Table 3, the results showed that DGMRL-MVC significantly outperformed DGCCA and DCCA on all datasets. In addition, the performance using the proposed discriminated representation was generally better than that using dynamic guided representation-MVC. In all, our model achieved comprehensive results because our model adds FDA-FISH and dynamic guided module that distills more separable intra-view and specific intrinsic features, resulting in a high-quality latent discriminated embedding.

**Table 3:**

**Comparison with latest methods with multiple views. The best results are highlighted in bold.**

*(RQ3) Is the discriminated fusion representation good on the clustering task?*

As shown in Table 4 and 5, the visualization is consistent with the clustering results. In this experiment, the *Silhouette_score* was also chosen as the evaluation indicator. Table 4 and 5 reveal the advantage of the proposed model when the dynamic guided and FDA-HSIC models were used simultaneously. The *Silhouette_score* for clustering was higher than that for the FISH-MML model. Besides, the *Silhouette_score* of the full view method was higher than that of the 2-view

382   method, which proved the necessity of the multi-view fusion representation. Our model has better
383   learning performance by using full views, and can effectively mine the features of multiple views.
384   The result empirically proves that clustering with full-view is more robust that with 2-view.
385   **Table 4:**
386   **Visualization of clustering result with FISH-MML on Pascal dataset. The best results are**
387   **highlighted in bold.**
388   **Table 5:**
389   **Visualization of clustering result with DGMRL-MVC (ours) on Pascal dataset. The best**
390   **results are highlighted in bold.**
391   *(RQ4) How does performance vary w.r.t. the parameter of n_clusters?*
392   Parameter analysis was conducted on the *Silhouette_score* parameter by varying the number
393   of clusters in [2,3,5,7,8,9,10] (Handwritten and Wikipedia datasets) and [2,5,8,10,13,15,18,20]
394   (Football and Pascal datasets). Fig. 2,3,4 and 5 plots the results in terms of *Silhouette_score* when
395   different numbers of clusters were used on the four datasets. The proposed DGMRL-MVC showed
396   the best performance on the other datasets. However, DGMRL-MVC slightly worse than the other
397   methods on the Handwritten dataset. Further, combined with the previous experimental results, we
398   analysis the reason of Figure 2 result. First, this dataset is different from other datasets, its views
399   are six types of features on same picture, i.e., Fourier coefficients of the character shapes, profile
400   correlations, Karhunen-love coefficients, pixel averages in $2 \times 3$ windows, Zernike moment,
401   morphological. Due to the large difference and weak correlation between the original data views,
402   our model weakens the original relationship of views. The value of the *Silhouette_score* reflects
403   the distance of samples in same cluster. In detail, although clustering evaluation metrics of our
404   model is higher than the baselines, the sample is closer to the boundary of the cluster in this kind
405   of dataset.
406   Figure 2: **Parameter analysis on *n_clusters* in terms of *Silhouette_score* on Handwritten**
407   **dataset.**
408   Figure 3: **Parameter analysis on *n_clusters* in terms of *Silhouette_score* on Wikipedia dataset.**
409   Figure 4: **Parameter analysis on *n_clusters* in terms of *Silhouette_score* on Football dataset.**
410   Figure 5: **Parameter analysis on *n_clusters* in terms of *Silhouette_score* on Pascal dataset.**
411   *(RQ5) Ablation experiment*
412   Finally, ablation experiments were used to further verify the effectiveness of the model. We
413   compared the differences of model learning performance under three steps respectively.
414   Specifically, Step-A is inter-intra representation, Step-B is fusion representation directly, Step-C is
415   dynamic routing and fusion representation. From the results of Figure 6, we can find that
416   performance of Step-B is lower than Step-C. The main reason is that fusion learning is only aimed
417   at the maximum correlation between views, lacking difference learning of views, which is not
418   conducive to the feature expression for clustering tasks. At the same time, the performance
419   difference between Step-B and Step-A is not big, and only some indicators are slightly higher than
420   Step-A, indicating that in the process of fusion learning, the specific intrinsic features learning with
421   dynamice routing plays the most obvious role. Our method result is higher than other sub-models,
422   which proves that on the basic of strengthening the learning view relationship, adding subspace
423   fusion characterization can effectively improve the clustering performance of model.
424   Figure 6: **Ablation experiments on four multi-view datasets.**

425 # Conclusions

426   This paper has proposed a novel framework called Dynamic Guided Metric Representation
427   Learning for Multi-View Clustering (DGMRL-MVC). The proposed method could cluster multi-

428 view data in a learned latent discriminated embedding space with multiple views by multi-step
429 enhanced representation. Within this framework, intra-view class separability and the complex
430 correlations among different views were explored using FDA-HSIC. Then the direction
431 information is added to enhance the single view representation by dynamic routing deep learning
432 mechanism. Finally, the common discriminated representation was obtained by generalized
433 canonical correlation analysis for multiple views and directly improved MVC performance.
434 Experiments on four multi-view datasets have demonstrated the effectiveness of the proposed
435 method for clustering tasks.
436      Multi-view data exits everywhere and there are still many challenging problems, such as data
437 alignment from heterogeneous sources (e.g. Wang, Zheng & L,2017), unified multi view fusion
438 model design and optimization (e.g. Tang et al,2020), and et al. In the future, we will do more
439 work on multi view data fusion.

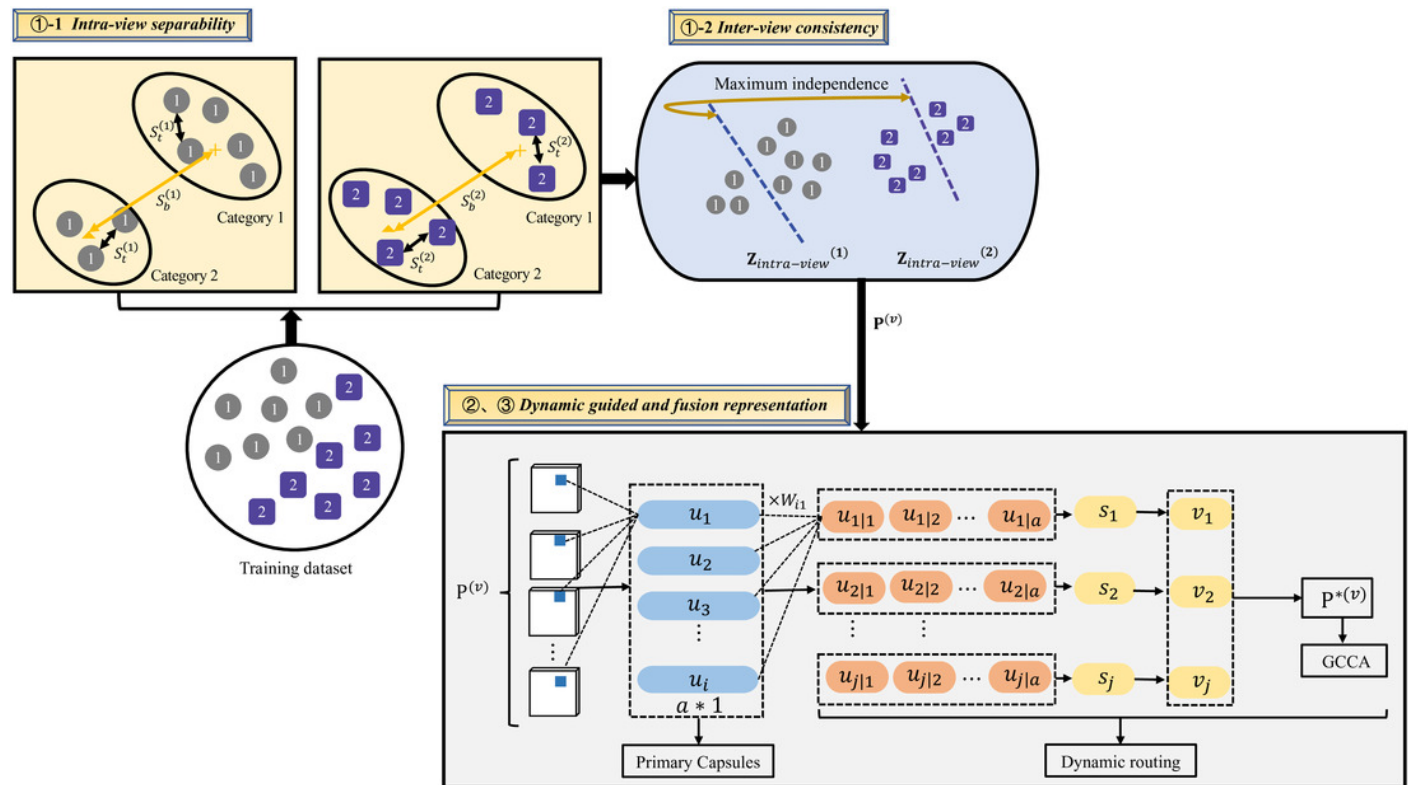
## References

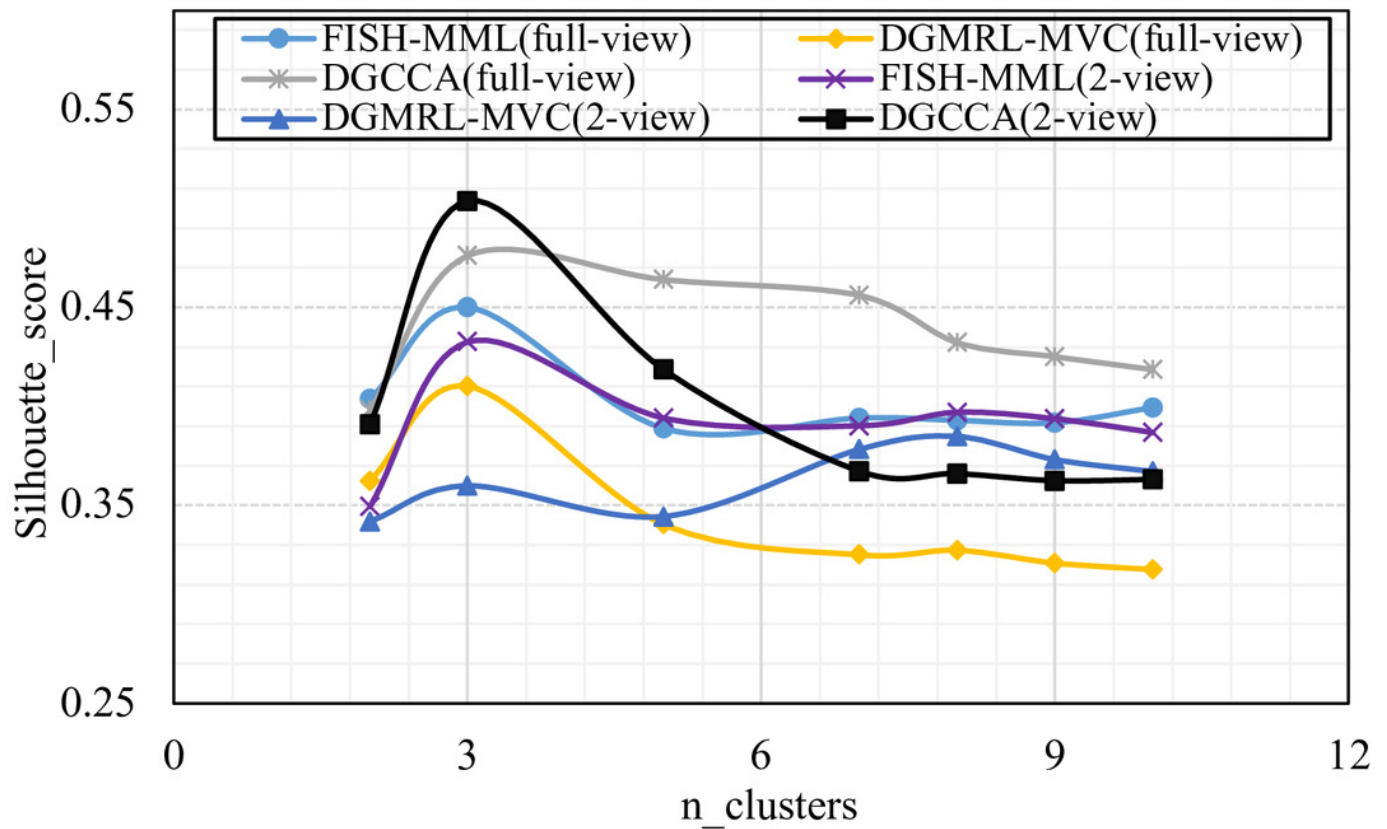
441 [1]Yang Y, Wang H. 2018. Multi-view clustering: A survey. journal of Big Data Min. Anal. *st*
442 1:83-107 DOI:10.26599/BDMA.2018.9020003.
443 [2]Wang H, Yang Y, Liu B. 2020. GMC: Graph-Based Multi-View Clustering. journal of IEEE
444 Transactions on Knowledge and Data Engineering *st* 32:1116-1129 DOI:10.1109/TKDE.2019.
445 2903810.
446 [3]Xia, Y, Yang, D, Yu Z. 2020. Uncertainty-aware multi-view co-training for semi-supervised
447 medical image segmentation and domain adaptation. Medical image analysis *st* 65,:101766
448 DOI:10.1016/j.media.2020.101766.
449 [4]Kang Z, Shi G, Huang S.2020. Multi-graph Fusion for Multi-view Spectral Clustering. journal
450 of Knowledge Based Systtems *st* 189 DOI:10.1016/j.knosys.2019.105102.
451 [5]Chen M, Huang L, Wang C. 2020. Multi-view Clustering in Latent Embedding Space. AAAI
452 DOI:10.1609/AAAI.V34I04.5756.
453 [6]Huang S, Kang Z, Xu, Z. 2020. Auto-weighted multi-view clustering via deep matrix
454 decomposition. journal of Pattern Recognit *st* 97 DOI:10.1016/j.patcog.2019.107015.
455 [7]Zhao H, Ding Z, Fu Y. 2017.Multi-View Clustering via Deep Matrix Factorization. AAAI.
456 [8]Haldorai A, Ramu A. 2020. Canonical Correlation Analysis Based Hyper Basis Feedforward
457 Neural Network Classification for Urban Sustainability. Neural Processing Letters:1-17
458 DOI:10.1007/s11063-020-10327-3.
459 [9]Hotelling H.1936.Relations between two sets of variates.Biometrika *st* 28:321-377
460 DOI:10.1007/978-1-4612-4380-9_14.
461 [10]Rasiwasia N, Mahajan DK, Mahadevan V. 2014. Cluster Canonical Correlation
462 Analysis. AISTATS.
463 [11]Blaschko MB, Lampert CH. 2008. Correlational spectral clustering. 2008 IEEE Conference
464 on Computer Vision and Pattern Recognition: 1-8 DOI:10.1109/CVPR.2008.4587353.
465 [12]Wang L,Li Y, Huang J.2017. Learning Two-Branch Neural Networks for Image-Text
466 Matching Tasks, journal of IEEE Transactions on Pattern Analysis and Machine Intelligence *st*
467 41:394-407 DOI:10.1109/TPAMI.2018.2797921.
468 [13]Andrew G, Arora R, Bilmes J.2013. Deep canonical correlation analysis. The 30th
469 International Conference on MachineLearning :1247–1255.
470 [14]A Benton,H Khayrallahand,B Gujral. 2017. Deep Generalized Canonical Correlation
471 Analysis.CoRR DOI:10.18653/v1/W19-4301.

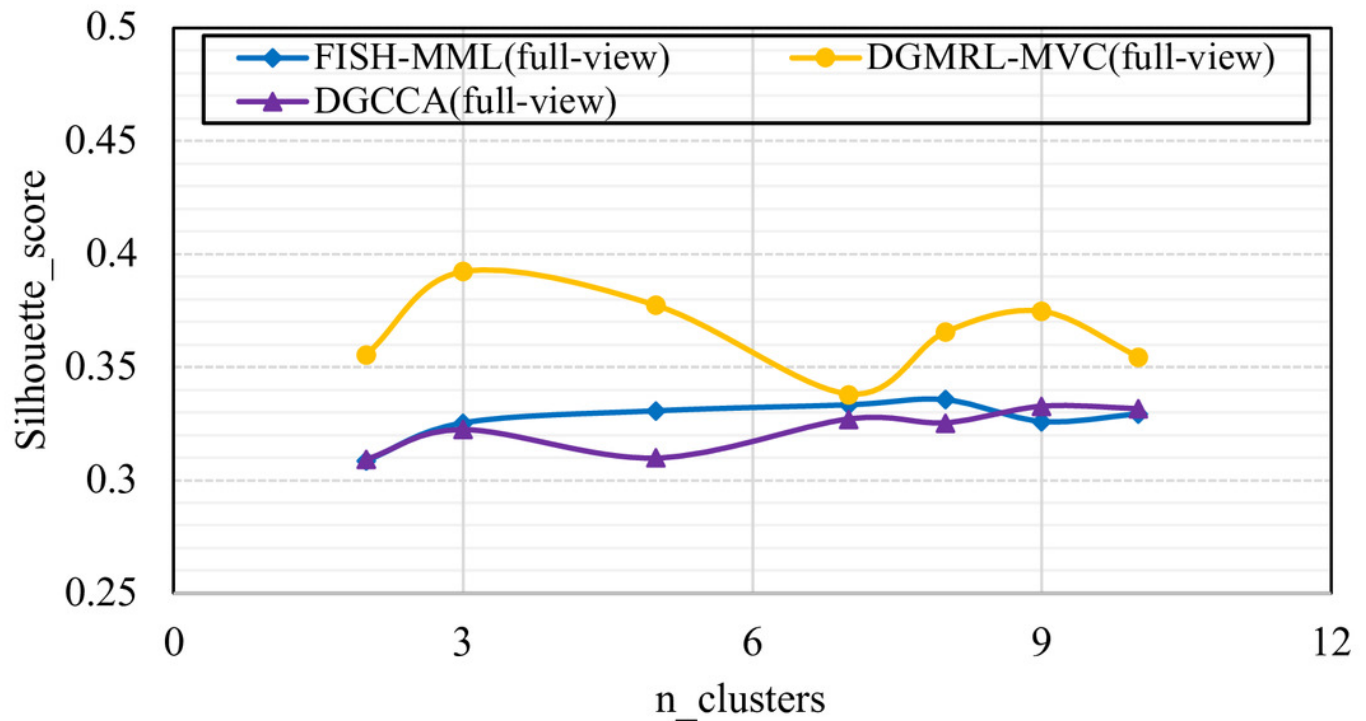
472 [15]Zheng T, Ge H, Li J.2020. Unsupervised multi-view representation learning with proximity
473 guided representation and generalized canonical correlation analysis. journal of Applied
474 Intelligence *st* 51:248-264 DOI:10.1007/s10489-020-01821-1.
475 [16]Zhu X, Zhang S, He W. 2019. One-Step Multi-View Spectral Clustering. Journal of IEEE
476 Transactions on Knowledge and Data Engineering, *st* 31:2022-2034
477 DOI:10.1109/TKDE.2018.2873378.
478 [17]Ye Y, Liu X, Yin J. 2016. Co-regularized kernel k-means for multi-view clustering. The 23rd
479 In7ternational Conference on Pattern Recognition: 1583-1588 DOI:10.1109/ICPR.2016.7899863.
480 [18]Chen M, Huang L, Wang C. 2020. Multi-View Clustering in Latent Embedding Space. AAAI
481 DOI:10.1609/AAAI.V34I04.5756.
482 [19]Wang Y, Lin X, Wu L. 2015. Robust Subspace Clustering for Multi-View Data by Exploiting
483 Correlation Consensus. journal of IEEE Transactions on Image Processing *st* 24: 3939-3949
484 DOI:10.1109/TIP.2015.2457339.
485 [20]Zhao H, Ding Z, Fu Y. 2017. Multi-View Clustering via Deep Matrix Factorization. AAAI.
486 [21]Kalayeh MM, Idrees H, Shah M. 2014. NMF-KNN: Image Annotation Using Weighted Multi-
487 view Non-negative Matrix Factorization. IEEE Conference on Computer Vision and Pattern
488 Recognition: 184-191 DOI:10.1109/CVPR.2014.31.
489 [22]Liu J, Wang Y, Lu R.2020. Multi-View Non-negative Matrix Factorization Discriminant
490 Learning via Cross Entropy Loss. 2020 Chinese Control And Decision Conference :3964-3971
491 DOI:10.1109/CCDC49329.2020.9163943.
492 [23]Sellami L, Alaya B.2021. SAMNET: Self-adaptative multi-kernel clustering algorithm for
493 urban VANETs. journal of Vehicular Communications *st* 29: 100332 DOI:10.1016/J.VEHCOM.
494 2021.100332.
495 [24]Lu H, Liu S, Wei H. 2020. Multi-kernel fuzzy clustering based on auto-encoder for fMRI
496 functional network. journal of Expert Systems with Appllications *st* 159: 113513 DOI:10.1016
497 /j.eswa.2020.113513.
498 [25]Nithya AA, Appathurai A, Venkatadri N. 2020. Kidney disease detection and segmentation
499 using artificial neural network and multi-kernel k-means clustering for ultrasound images. journal
500 of Measurement *st* 149:106952 DOI:10.1016/J.MEASUREMENT.2019.106952.
501 [26]Cui Y, Fern XZ, Dy JG. 2007. Non-redundant Multi-view Clustering via
502 Orthogonalization. Seventh IEEE International Conference on Data Mining: 133-142
503 DOI:10.1109/ICDM.2007.94.
504 [27]Liu X, Ji S, Glänzel W. 2013. Multiview Partitioning via Tensor Methods. journal of IEEE
505 Transactions on Knowledge and Data Engineering *st* 25: 1056-1069 DOI:10.1109/TKDE.2012.95.
506 [28]Monney A, Zhan Y, Zhen J. 2020. A multi-kernel method of measuring adaptive similarity for
507 spectral clustering. journal of Expert Systems with Appllications *st* 159:113570
508 DOI:10.1016/j.eswa.2020.113570.
509 [29]Xu J, Han J, Nie F. 2016. Discriminatively Embedded K-Means for Multi-view Clustering.
510 IEEE Conference on Computer Vision and Pattern Recognition :5356-5364
511 DOI:10.1109/CVPR.2016.578.
512 [30]Nie F,Li J,Li X.2017. Self-weighted multiview clustering with multiple graphs. The Twenty-
513 Sixth International Joint Conference on Artificial Intelligence:2564-2570 DOI:10.24963
514 /ijcai.2017/357.
515 [31]Chaudhuri K, Kakade SM, Livescu K.2009. Multi-view clustering via canonical correlation
516 analysis. ICML '09 DOI:10.1145/1553374.1553391.

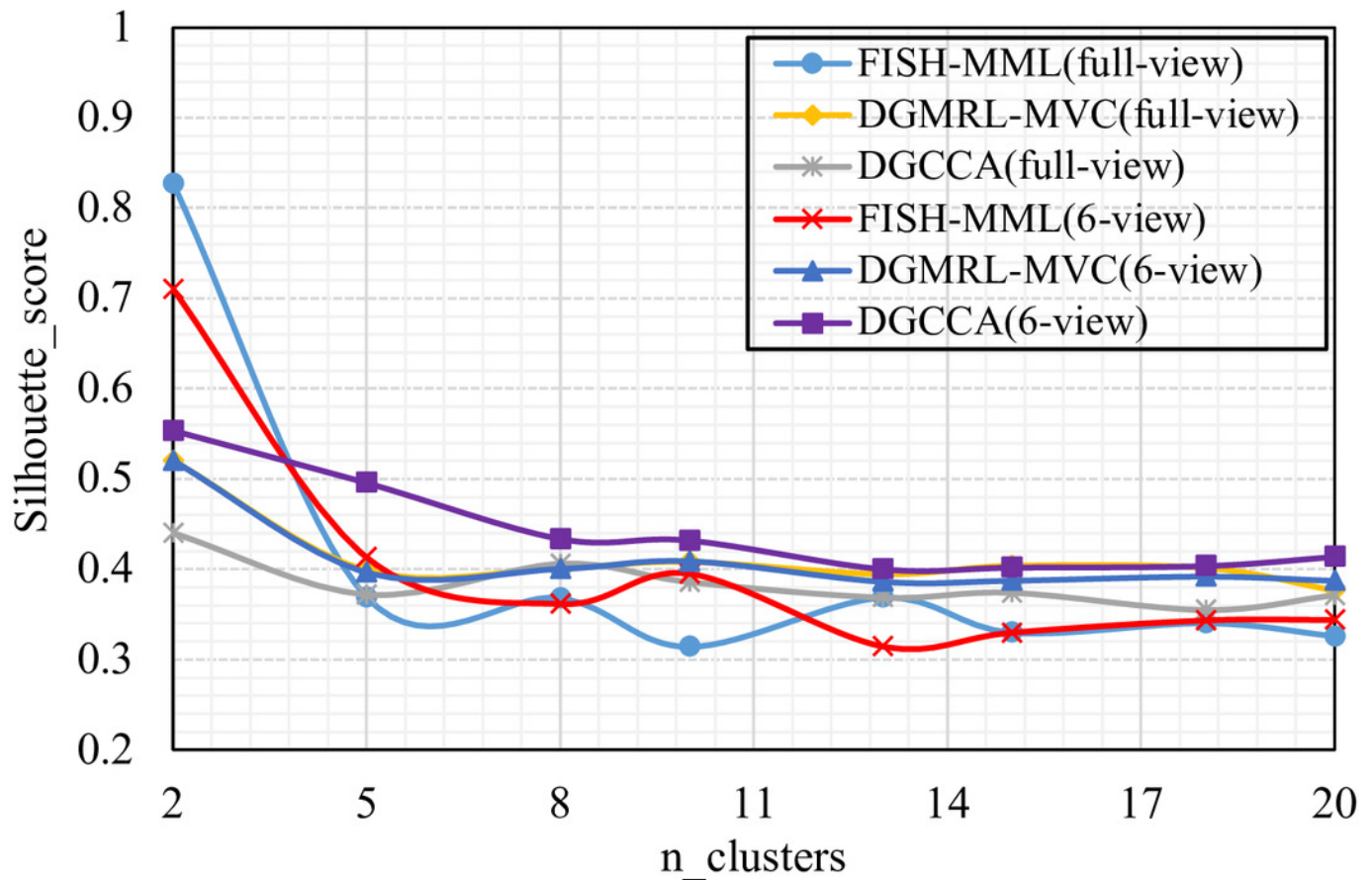
517    [32]Yan F,Mikolajczyk K.2015. Deep correlation for matching images and text. CVPR:3441–
518    3450 DOI:10.1109/CVPR.2015.7298966.
519    [33]Li J,Yong H,Zhang B. 2018.A Probabilistic Hierarchical Model for Multi-View and Multi-
520    Feature Classification. AAAI.
521    [34]Benton A,Khayrallah H,Gujral B.2017. Deep Generalized Canonical Correlation
522    Analysis.CoRR DOI:10.18653/v1/W19-4301.
523    [35]Zhang C, Liu Y, Liu Y.2018. FISH-MML: Fisher-HSIC Multi-View Metric Learning. IJCAI
524    DOI:10.24963/ijcai.2018/424.
525    [36]Gretton A, Bousquet O, Smola AJ. 2005. Measuring Statistical Dependence with Hilbert-
526    Schmidt Norms. journla of ALT $st$ 16:63-78 DOI:10.1007/11564089_7.
527    [37]Sabour S, Frosst N, & Hinton GE. 2017. Dynamic Routing Between Capsules. ArXiv:
528    Computer Vision and Pattern Recognition.abs/1710.09829.
529    [38]Cyrus R, Peter Y, Micah H.2010.Collecting Image Annotations Using Amazon's Mechanical
530    Turk. The NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's
531    Mechanical Turk.
532    [39]Chandar S, Khapra MM, Larochelle H. 2015.Correlational neural networks. CoRR
533    DOI:10.1162/NECO_a_00801.
534    [40]Wang L, Zheng T, Li M. 2017. A Survey of Key Technologies and Applications on
535    Heterogeneous Network Alignment in Social Network Big Data. Journal of Taiyuan University of
536    Technology $st$ 48(3): 453-457 DOI:10.16355/j.cnki.issn1007-9432tyut.2017.03.024.
537    [41]Tang X, Zhang Q, Dai X, Zou Y. 2020. Neural membrane mutual coupling characterisation
538    using entropy-based iterative learning identification. journal of IEEE Access $st$ 8:205231-205243
539    DOI:10.1109/ACCESS.2020.3037816.
540    [42]Cyrus R, Peter Y, Micah Hodosh.2010. Collecting Image Annotations Using Amazon's
541    Mechanical Turk. NAACL HLT 2010 Workshop on Creating Speech and Language Data with
542    Amazon's Mechanical Turk.
543    [43]Greene D, & Cunningham, P. 2013. Producing a unified graph representation from multiple
544    social network views. WebSci.  DOI:10.1145/2464464.2464471.

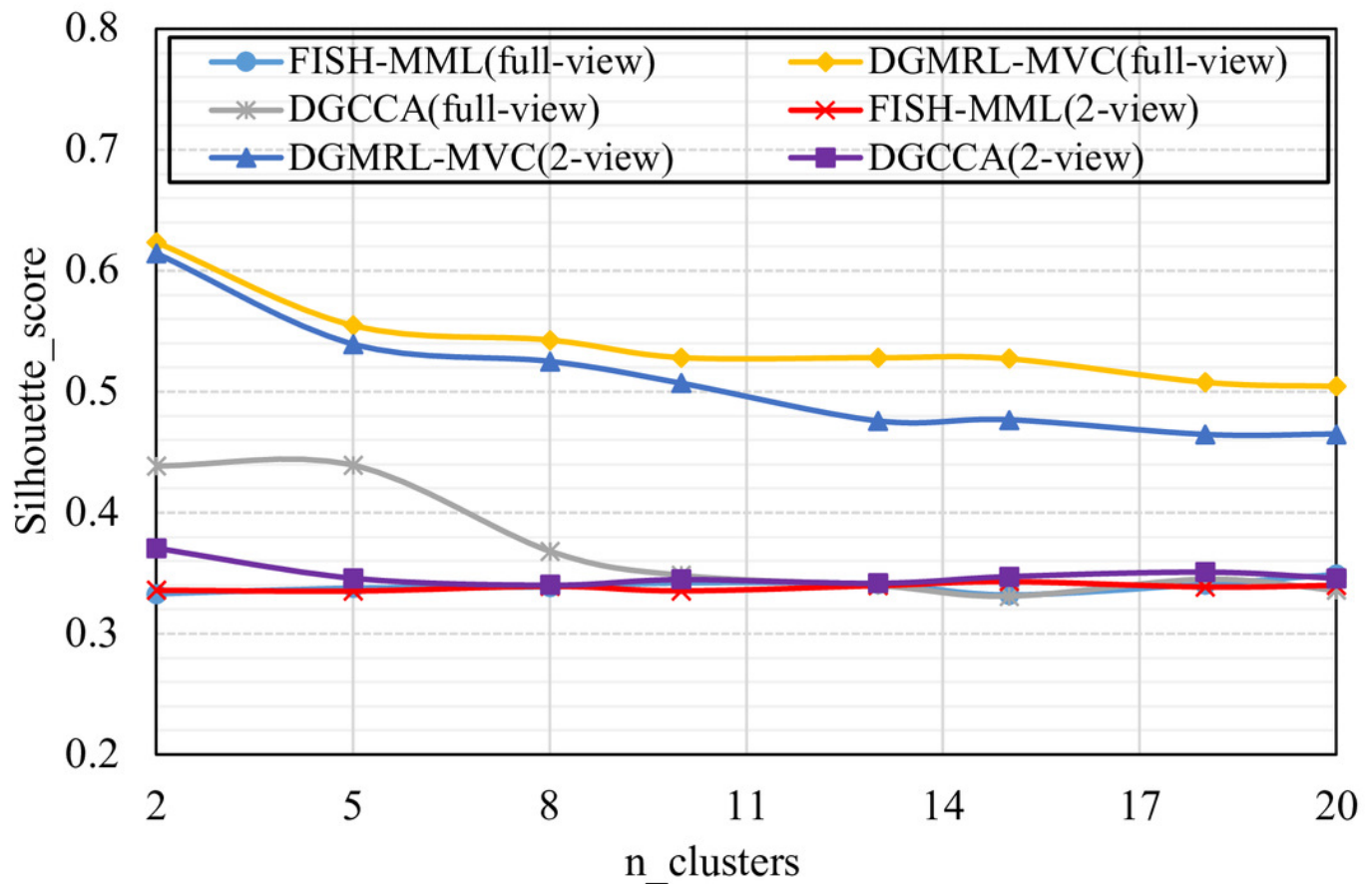
# Figure 1

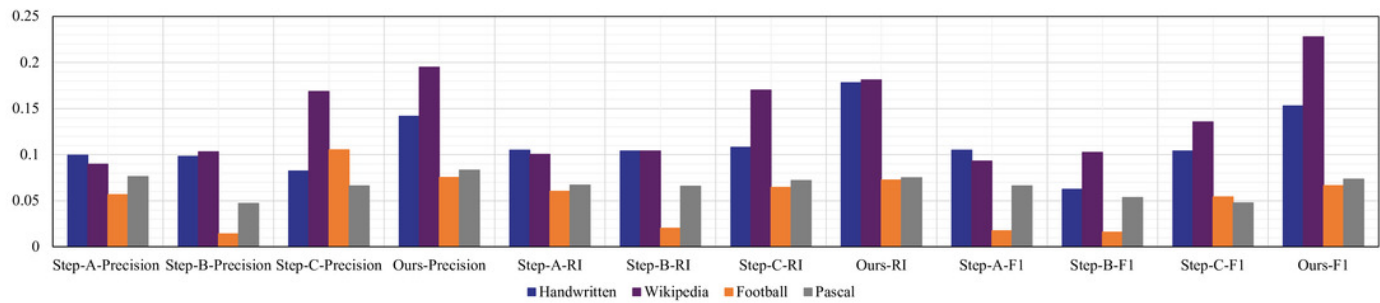
Illustration of the proposed DGMRL-MVC model

# Figure 2

Parameter analysis on *n_clusters* in terms of *Silhouette_score* on Handwritten dataset

# Figure 3

Parameter analysis on *n_clusters* in terms of *Silhouette_score* on Wikipedia dataset

# Figure 4

Parameter analysis on *n_clusters* in terms of *Silhouette_score* on Football dataset

# Figure 5

Parameter analysis on *n_clusters* in terms of *Silhouette_score* on Pascal dataset

# Figure 6

Ablation experiments on four multi-view datasets.

The differences among baselines

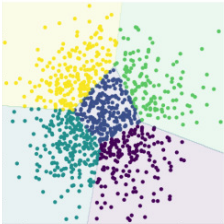
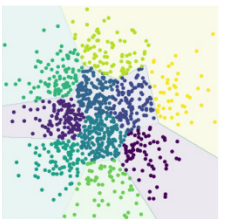
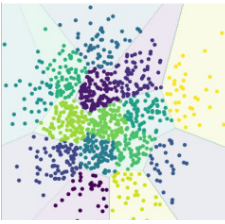
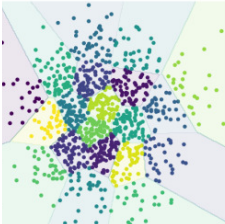
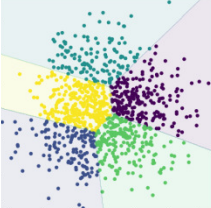
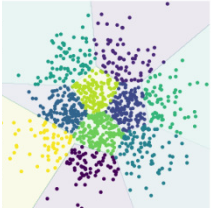
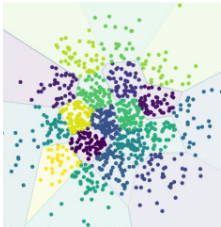
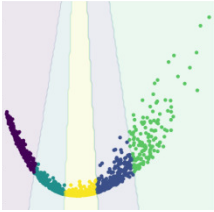
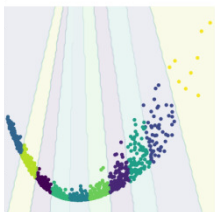
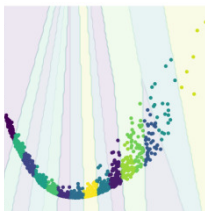
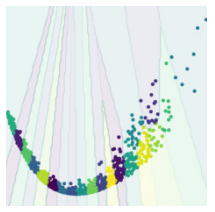
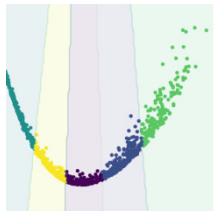
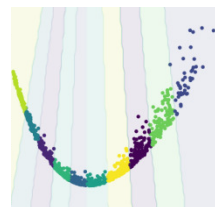
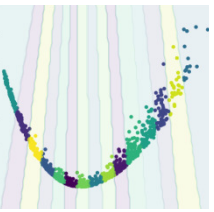
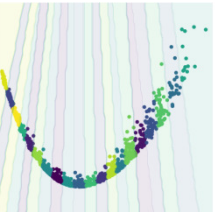
| Method | Type | Dynamic guided representation | FDA-HSIC | Optimization |
|---|---|---|---|---|
| DCCA | Deep, 2-view | × | × | $\left(\theta_1^*,\theta_2^*\right)=arg\max\limits_{(\theta_1,\theta_2)}\left(f_1(X_1;\theta_1),f_2(X_2;\theta_2)\right)$ |
| DGCCA | Deep, multi-view | × | × | $\underset{U_j\in\mathbb{R}^{d_j\times r},\,G\in\mathbb{R}^{r\times N}}{\text{Minimize}}\sum\limits_{j=1}^{J}\left\|\mathbf{G}-\mathbf{U}_j^{\mathbf{T}}f_j(X_j)\right\|_F^2$ |
| FISH-MML | Deep, multi-view | × | √ | $\underset{\{M^{(v)}\}_{v=1}^{V}}{max}\,S\left(\{\mathbf{M}^{(v)}\}_{v=1}^{V}\right)+\lambda C\left(\{\mathbf{M}^{(v)}\}_{v=1}^{V}\right)$ |
| Dynamic guided representation-MVC | Deep, multi-view | √ | × | $\underset{U_j\in\mathbb{R}^{d_j\times r},\,G\in\mathbb{R}^{r\times N}}{\text{Minimize}}\sum\limits_{j=1}^{J}\left\|\mathbf{G}-\mathbf{U}_j^{T}\mathbf{O}_{\mathbf{j}}^*\right\|_F^2$ |
| DGMRL-MVC | Deep, multi-view | √ | √ | $\underset{U_v\in\mathbb{R}^{d_v\times r},\,G\in\mathbb{R}^{r\times N}}{\text{minimize}}\sum\limits_{v=1}^{V}\left\|\mathbf{G}-\mathbf{U}_{\mathbf{v}}^{\mathbf{T}}\mathbf{P}^{*(v)}\right\|_{\mathrm{F}}^2$ |

Improvement performance of dynamic guided representation on different views. The best results are highlighted in bold.

| Dataset | Method | Number of views | Precision | Recall | F1 | RI | NMI |
|---------|--------|-----------------|-----------|--------|-----|-----|-----|
| Handwritten | FISH-MML | 6 (full views) | 0.09961 | 0.11099 | 0.10493 | 0.11100 | 0.09643 |
| | DGMRL-MVC | | **0.14166** | **0.17800** | **0.15315** | **0.17800** | **0.11521** |
| | FISH-MML | 2 views | 0.17843 | 0.20600 | 0.18586 | 0.20600 | 0.21906 |
| | DGMRL-MVC | | **0.25699** | **0.22100** | **0.22521** | **0.22100** | **0.32698** |
| | FISH-MML | 3 views | 0.09092 | 0.10100 | 0.09569 | 0.10100 | 0.25037 |
| | DGMRL-MVC | | **0.17813** | **0.14300** | **0.13452** | **0.14300** | **0.37442** |
| Wikipedia | FISH-MML | 2 (full views) | 0.08957 | 0.10149 | 0.09313 | 0.10032 | 0.51780 |
| | DGMRL-MVC | | **0.10491** | **0.12321** | **0.10398** | **0.13438** | **0.53230** |
| Football | FISH-MML | 9 (full views) | 0.05679 | 0.05885 | 0.01761 | 0.06048 | 0.22956 |
| | DGMRL-MVC | | **0.05900** | **0.06248** | **0.05575** | **0.06452** | **0.28367** |
| | FISH-MML | 6 views | **0.11914** | 0.04903 | 0.03745 | 0.05645 | **0.28422** |
| | DGMRL-MVC | | 0.05619 | **0.06506** | **0.05848** | **0.06452** | 0.27343 |
| Pascal | FISH-MML | 6 (full views) | 0.07616 | 0.06700 | 0.06606 | 0.06700 | 0.23351 |
| | DGMRL-MVC | | **0.08315** | **0.07500** | **0.07342** | **0.07500** | **0.48847** |
| | FISH-MML | 2 views | 0.06266 | 0.05600 | 0.05094 | 0.05600 | 0.18011 |
| | DGMRL-MVC | | **0.08313** | **0.06800** | **0.06448** | **0.06800** | **0.45506** |

Comparison with latest methods with multiple views. The best results are highlighted in bold.

| Dataset | Method | Number of views | Precision | Recall | F1 | RI | NMI |
|---------|--------|-----------------|-----------|--------|-----|-----|-----|
| Handwritten | DCCA | 6 (full views) | 0.10111 | 0.10000 | 0.10047 | 0.10000 | **0.57354** |
| | DGCCA | | 0.08222 | 0.05400 | 0.06244 | 0.10800 | 0.54150 |
| | FISH-MML | | 0.09961 | 0.11099 | 0.10493 | 0.11100 | 0.09643 |
| | Dynamic guided representation-MVC | | 0.09835 | 0.11650 | 0.10402 | 0.11650 | 0.52245 |
| | DGMRL-MVC | | **0.14166** | **0.17800** | **0.15315** | **0.17800** | 0.11521 |
| Wikipedia | DCCA | 2 (full views) | 0.09245 | 0.08467 | 0.08289 | 0.09090 | 0.50819 |
| | DGCCA | | 0.10329 | 0.10567 | 0.10267 | 0.10400 | 0.54222 |
| | FISH-MML | | 0.08957 | 0.10149 | 0.09313 | 0.10032 | 0.51780 |
| | Dynamic guided representation-MVC | | 0.16876 | 0.14224 | 0.13550 | 0.17027 | 0.59470 |
| | DGMRL-MVC | | **0.19480** | **0.18393** | **0.22795** | **0.18110** | **0.59851** |
| Football | DCCA | 9 (full views) | 0.05625 | 0.04909 | 0.05200 | 0.05241 | 0.21037 |
| | DGCCA | | 0.01404 | 0.01885 | 0.01603 | 0.02016 | 0.21256 |
| | FISH-MML | | 0.05679 | 0.05885 | 0.01761 | 0.06048 | 0.22956 |
| | Dynamic guided representation-MVC | | **0.10516** | 0.06449 | 0.05430 | 0.06451 | 0.26470 |
| | DGMRL-MVC | | 0.07540 | **0.06867** | **0.06647** | **0.07258** | **0.28367** |
| Pascal | DCCA | 6 (full views) | 0.02123 | 0.04200 | 0.02814 | 0.04200 | 0.03363 |
| | DGCCA | | 0.04716 | 0.06600 | 0.05351 | 0.06600 | 0.09261 |
| | FISH-MML | | 0.07616 | 0.06700 | 0.06606 | 0.06700 | 0.23351 |
| | Dynamic guided representation-MVC | | 0.06619 | 0.07200 | 0.04782 | 0.07200 | 0.15842 |
| | DGMRL-MVC | | **0.08315** | **0.07500** | **0.07342** | **0.07500** | **0.48847** |

Visualization of clustering result with FISH-MML on Pascal dataset. The best results are highlighted in bold.

| | | | |
|---|---|---|---|
|  |  |  |  |
| *n_clusters*=5 | *n_clusters*=10 | *n_clusters*=15 | *n_clusters*=20 |
| 2-view | 2-view | 2-view | 2-view |
| *Silhouette_score*= 0.31934 | *Silhouette_score*=0.34125 | *Silhouette_score*= 0.33213 | *Silhouette_score*= 0.32979 |
|  |  |  |  |
| *n_clusters*=5 | *n_clusters*=10 | *n_clusters*=15 | *n_clusters*=20 |
| Full-view | Full-view | Full-view | Full-view |
| ***Silhouette_score*= 0.31940** | ***Silhouette_score*=0.33300** | ***Silhouette_score*=0.32456** | ***Silhouette_score*= 0.34295** |

Visualization of clustering result with DGMRL-MVC (ours) on Pascal dataset. The best results are highlighted in bold

| | | | |
|---|---|---|---|
| *n_clusters*=5 | *n_clusters*=10 | *n_clusters*=15 | *n_clusters*=20 |
| 2-view | 2-view | 2-view | 2-view |
| *Silhouette_score*=0.53981 | *Silhouette_score*=0.51019 | *Silhouette_score*=0.47236 | *Silhouette_score*=0.46940 |
| *n_clusters*=5 | *n_clusters*=10 | *n_clusters*=15 | *n_clusters*=20 |
| Full-view | Full-view | Full-view | Full-view |
| ***Silhouette_score*=0.55487** | ***Silhouette_score*=0.53538** | ***Silhouette_score*=0.52072** | ***Silhouette_score*=0.51451** |