Gromov-Wasserstein Multi-modal Alignment and Clustering

Fengjiao Gong*
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
gongfengjiao2021@ruc.edu.cn

Yuzhou Nie*
School of Statistics
Renmin University of China
Beijing, China
nieyuzhou@ruc.edu.cn

Hongteng Xu†
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
hongtengxu@ruc.edu.cn

ABSTRACT

Multi-modal clustering aims at finding a clustering structure shared by the data of different modalities in an unsupervised way. Currently, solving this problem often relies on two assumptions: *i*) the multi-modal data own the same latent distribution, and *ii*) the observed multi-modal data are well-aligned and without any missing modalities. Unfortunately, these two assumptions are often questionable in practice and thus limit the feasibility of many multi-modal clustering methods. In this work, we develop a new multi-modal clustering method based on the Gromovization of optimal transport distance, which relaxes the dependence on the above two assumptions. In particular, given the data of different modalities, whose correspondence is unknown, our method learns the Gromov-Wasserstein (GW) barycenter of their kernel matrices. Driven by the modularity maximization principle, the GW barycenter helps to explore the clustering structure shared by different modalities. Moreover, the GW barycenter is associated with the GW distances between the different modalities to the clusters, and the optimal transport plans corresponding to the GW distances help to achieve the alignment and the clustering of the multi-modal data jointly. Experimental results show that our method outperforms state-of-the-art multi-modal clustering methods, especially when the data are (partially or completely) unaligned. The code is available at https://github.com/rucnyz/GWMAC.

CCS CONCEPTS

• **Computing methodologies** → **Cluster analysis**; **Spectral methods**.

KEYWORDS

Multi-modal clustering, Gromov-Wasserstein barycenter, kernel fusion, optimal transport, data alignment

ACM Reference Format:
Fengjiao Gong, Yuzhou Nie, and Hongteng Xu. 2022. Gromov-Wasserstein Multi-modal Alignment and Clustering. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3511808.3557339

*Equal contribution.
†Corresponding author: Hongteng Xu (hongtengxu@ruc.edu.cn)

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557339

1 INTRODUCTION

Many real-world machine learning tasks involve exploring the clustering structures of the unlabeled data collected from different resources and in different formats, which leads to the so-called "multi-modal clustering" problem. For precise medicine, the analysis of a disease depends on the clustering of patients' electronic health records (EHRs) that contain heterogeneous data modalities like clinical notes (texts), lab results (tablets), and medical images [71, 75]. For machine translation and multi-linguistic text classification, the words of different languages should be matched and clustered according to their semantics [19, 61]. Besides the above two examples, multi-modal clustering is also significant for other practical applications, such as computer vision [32, 62] and cross-modal data generation [43, 44].

To achieve multi-modal clustering, many methods have been proposed, which can be coarsely categorized into two strategies: co-regularization [26] and kernel fusion [15]. In particular, co-regularization aims at learning latent representations shared by different modalities and achieves clustering in the latent space accordingly. The commonly-used latent representation methods include canonical correlation analysis [5, 50], low-rank approximation [11, 16], non-negative matrix factorization [30, 76], and their neural network-based variants [3, 28, 57, 78]. Kernel fusion, on the other hand, aims at leveraging the relational information of different modalities jointly and fusing the information for clustering. Typically, for the samples of each modality, their relational information can be the graph structure [27, 54], the distance matrix [34, 39], and so on. Based on the relational information, we can construct and fuse the kernel matrices of different modalities [10, 49] and then apply spectral clustering [38, 80] or k-means [12, 31] to achieve multi-modal clustering.

Both of the above two strategies are dependent on two assumptions: *i*) the multi-modal data own the same latent distribution or clustering structure, and *ii*) the observed multi-modal data are well-aligned and without any missing modalities. However, these two assumptions are often questionable in practice. In particular, different modalities may contain complementary information, and some modalities can even be useless in some tasks. Therefore, their latent distributions can be different. Additionally, the real-world multi-modal data can be collected from different resources in different trials, and the samples of different modalities can be independent and unaligned (i.e., the correspondence between the samples of different modalities is unknown). For example, the EHRs

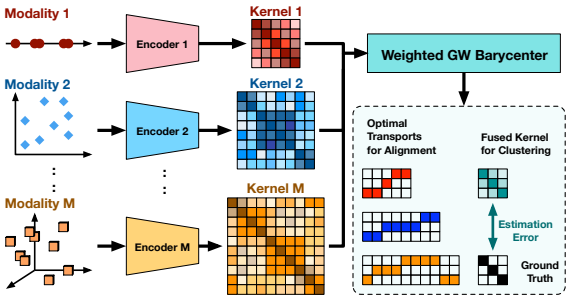


Figure 1: An illustration of the proposed GWMAC method.

for disease analysis can be collected from the patients in different hospitals. Each patient's EHR may just contain the information of partial modalities because of the lack of medical resources or the restrictions of insurance coverage, and some modalities (e.g., lab results, genetic tests) may be more important than the others (e.g., drug records) for disease diagnosis and patient clustering. Faced with such practical and challenging multi-modal data, most existing multi-modal clustering methods either lead to sub-optimal performance or become inapplicable.

To relax the dependency on the above two assumptions, we propose a novel Gromov-Wasserstein multi-modal alignment and clustering (GWMAC) method. As illustrated in Figure 1, our GWMAC method neither requires the observed multi-modal data to be well-aligned nor restricts different modalities to share the same latent distribution/structure. For the samples of each modality, it derives their latent representations through a learnable encoder and constructs a kernel matrix. Without the correspondence between different modalities' samples, our method fuses the kernel matrices by solving a weighted Gromov-Wasserstein (GW) barycenter problem [42, 64]. The barycenter works as a fused kernel matrix, whose GW distances to the kernel matrices are minimized. Different from the work in [64], which learns the latent representation of graphs based on a factorization model and applies K-means accordingly, our method is built based on the modularity maximization principle and thus can indicate the clustering structure of the data directly. Additionally, solving the GW barycenter problem provides us with a set of optimal transport plans to align the samples of different modalities to the clusters.

Different from most existing multi-modal clustering methods, our GWMAC achieves the alignment and the clustering of multi-modal data jointly, which is applicable even if the multi-modal data are **totally** unaligned. Additionally, our method does not require the modalities to share the same latent distributions, because it is based on the modalities' kernel matrices rather than using their latent codes directly. Note that, the weights associated with different kernel matrices are learnable, so instead of treating different modalities evenly, our GWMAC method learns the significance of different modalities, which is robust to those noise and useless modalities. We demonstrate the feasibility and the superiority of our method on several representative datasets. Experimental results show that our GWMAC performs the best or beyond average in well-aligned setting, and significantly outperforms state-of-the-art methods in both partially-aligned and totally-unaligned settings.

2 RELATED WORK

2.1 Multi-modal clustering

Most existing multi-modal clustering methods depend on either the co-regularization strategy or the kernel fusion strategy. For the co-regularization strategy, the canonical correlation analysis (CCA) has been widely used. CCA maps different modalities to the same latent space and maximizes the correlation of their latent codes accordingly [50]. The early CCA-based methods apply linear mapping functions [5]. With the development of deep learning, the mapping functions can be parameterized via deep neural networks (i.e., encoders) [58, 74, 81]. Besides maximizing the correlation, other regularizers can be applied when learning the encoders, e.g., considering the reconstruction errors of different modules in an auto-encoding framework [41, 73] and introducing adversarial regularizers for the latent codes [28, 57]. Besides the above CCA-based methods, some other methods impose structural constraints on the latent codes of different modalities, e.g., orthogonal constraint [68], low-rank structure [11, 16, 79], non-negativeness constraint [30, 53].

The kernel fusion strategy is commonly applied to the multi-modal data with significant structural information. For example, in many applications, the samples of each modality may own a graph structure [20, 54], and thus, the relation or the similarity between arbitrary two samples can be represented via various kernel matrices [18, 34, 49]. The valid kernel matrix, which is derived by fusing the kernel matrices of different modalities, indicates their shared clustering structure. Accordingly, the multi-modal clustering is achieved via applying spectral clustering [18, 25, 80] or k-means [12, 17, 31] to the fused kernel matrix. In general, the kernel matrices are fused in an additive way, where the kernels and their weights can be fixed [10] or learnable [55]. The work in [55] first partitions each view by kernel k-means, then maximizes the alignment between the weighted partitions so as to reduce the computation complexity. Recently, the relations among the samples can be adjusted or optimized during training as well, which leads to the combination of the kernel fusion methods with other graph-based learning frameworks [27, 40, 54, 77]. Note that the kernel fusion strategy does not map different modalities to the same latent space explicitly. Such flexibility motivates us to implement our GWMAC method based on kernels.

2.2 Multi-modal alignment

These above multi-modal clustering methods require that the samples of different modalities are well-aligned, i.e., the correspondence between the samples is known. To relax this strict constraint, some attempts have been made to achieve multi-modal clustering in more challenging scenarios. For example, methods in [13, 59, 60, 63, 69] achieve multi-modal clustering based on incomplete multi-modal data (i.e., for some multi-modal samples, a part of their modalities are unobserved). Besides the incompleteness, the unbalance issue (i.e., some modalities own few samples while the others have lots of observations) [14] and the inconsistency issue (i.e., the modalities contains complementary even inconsistent information) [7] are also considered. These issues can be viewed as the special cases of the incomplete problem. As a result, we can also deal with these issues by aligning the samples of different modalities, but it requires us to consider the significance of the modalities at the same time.

Essentially, these above methods aim at estimating the correspondence of the samples across different modalities, which leads to the multi-modal alignment problem. However, most existing multi-modal alignment methods require at least a part of well-aligned data [22, 24, 29]. Although some recent work has made efforts to align the totally-unaligned multi-modal data [4, 66, 70], their performance suffers from the identifiability issue because the alignment problem itself is NP-hard. As a result, their alignment results are normally inconsistent when the number of modalities is larger than two and are sensitive to the noise of data.

2.3 Optimal transport-based machine learning

Recently, the optimal transport theory [51] shows the potentials to various machine learning tasks, such as distribution matching [2], generative modeling [1, 48], shape comparison [35], graph analysis [6], and so on. Some recent work demonstrates that the well-known Wasserstein distance [23] and its Gromovized variants (i.e., the Gromov-Wasserstein distance) [35] can be used to achieve data alignment and clustering [8, 66]. The optimal transport plan associated with these distances helps to estimate the correspondence between samples (or between the samples and the clusters). Recently, for the structured data like point clouds and graphs, the Gromov-Wasserstein distance [46] has been proven useful for the alignment problems, such as graph matching [47] and point cloud registration [42]. More recently, its capability of clustering is explored by the work in [33, 65] and is demonstrated in [6]. The proposed GWMAC method extends the GW spectral method in [6] to multi-modal scenarios.

3 PROPOSED METHOD

Suppose that we observe some data of M modalities, i.e., $\{\boldsymbol{X}_m \in \mathbb{R}^{N_m \times D_m}\}_{m=1}^{M}$ and each $\boldsymbol{X}_m = \{\boldsymbol{x}_{m,i} \in \mathbb{R}^{D_m}\}_{i=1}^{N_m}$ contains N_m D_m-dimensional samples and corresponds to a specific modality. Different from the typical scenarios considered by most existing methods, the multi-modal data are unaligned, i.e., the correspondence between arbitrary two samples of different modalities is unknown, and generally, $N_m \neq N_{m'}$ for $m \neq m'$. Given such unaligned multi-modal data, we aim to align the samples across different modalities and explore the clustering structure shared by the modalities.

Obviously, the alignment and the clustering of the multi-modal data are highly correlated so that each problem has impact on the other one. In this study, we model these two problems jointly in a kernel fusion framework and show that this learning task can be solved efficiently with the help of optimal transport techniques.

3.1 A joint alignment-clustering framework

Typically, when the multi-modal data are well aligned, i.e., $\boldsymbol{X} = [\boldsymbol{X}_1, ..., \boldsymbol{X}_M] \in \mathbb{R}^{N \times (D_1 + ... + D_M)}$, most existing kernel fusion-based clustering methods [25, 49, 54, 80] can be formulated as follows:

$$\begin{aligned} &\max_{\boldsymbol{G} \in \Omega, \boldsymbol{\theta}} \operatorname{tr}(\boldsymbol{G}^T \bar{\boldsymbol{K}}(\boldsymbol{X}; \boldsymbol{\theta}) \boldsymbol{G}) \\ &s.t.\ \bar{\boldsymbol{K}}(\boldsymbol{X}; \boldsymbol{\theta}) = \sum\nolimits_{m=1}^{M} \alpha_m \boldsymbol{K}_m(\boldsymbol{X}_m; \boldsymbol{\theta}_m), \end{aligned} \tag{1}$$

where $\operatorname{tr}(\cdot)$ represents the trace of matrix. $\bar{\boldsymbol{K}}(\boldsymbol{X}; \boldsymbol{\theta}) \in \mathbb{R}^{N \times N}$ is the fused kernel matrix parameterized by the model parameter $\boldsymbol{\theta}$. It is constructed as the sum of the weighted kernel matrices of different modalities, i.e., $\{\boldsymbol{K}_m(\boldsymbol{X}_m; \boldsymbol{\theta}_m)\}_{m=1}^{M}$,[1] and the weight vector $\boldsymbol{\alpha} = [\alpha_m]$ is in the $(M-1)$-simplex, i.e., $\boldsymbol{\alpha} \in \Delta^{M-1}$. The weight α_m can be interpreted as the significance of the m-th modality. Accordingly, the model parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M, \boldsymbol{\alpha}\}$ include the parameters for each modality-specific kernel and the weights of the modalities. $\boldsymbol{G} \in \Omega$ is an indicator matrix that indicates the clustering structure. When the feasible domain $\Omega = \{\boldsymbol{G} \in \mathbb{R}^{N \times d} | \boldsymbol{G}^T \boldsymbol{G} = \boldsymbol{I}_d\}$, the objective function in (1) corresponds to the spectral clustering. The objective function in (1) becomes kernel K-means [12, 17, 31] when $\Omega = \{\boldsymbol{G} \in \{0,1\}^{N \times d} | \boldsymbol{G}\boldsymbol{1}_d = \boldsymbol{1}_N\}$, where d is the desired number of clusters.

In our task, however, the multi-modal data are unaligned. Therefore, we need to consider the alignment of the data before fusing their kernel matrices. A naïve solution to this problem is first aligning the kernel matrices pairwisely and then applying the kernel fusion-based clustering. In particular, given two kernel matrices $\boldsymbol{K}_m \in \mathbb{R}^{N_m \times N_m}$ and $\boldsymbol{K}_{m'} \in \mathbb{R}^{N_{m'} \times N_{m'}}$, where $N_m \geq N_{m'}$, their alignment corresponds to solving the following quadratic assignment programming (QAP) problem:

$$\boldsymbol{T}_{m,m'} = \arg\max\nolimits_{\boldsymbol{T} \in P(\boldsymbol{1}_{N_m}, \boldsymbol{1}_{N_{m'}})} \operatorname{trace}(\boldsymbol{K}_{m'}^T \boldsymbol{T}^T \boldsymbol{K}_m \boldsymbol{T}), \tag{2}$$

where $\boldsymbol{T}_{m,m'}$ is the optimal alignment matrix that matches $\boldsymbol{K}_m$ with $\boldsymbol{K}_{m'}$. $P(\boldsymbol{1}_{N_m}, \boldsymbol{1}_{N_{m'}}) = \{\boldsymbol{T} \in \{0,1\}^{N_m \times N_{m'}} | \boldsymbol{T}\boldsymbol{1}_{N_{m'}} \leq \boldsymbol{1}_{N_m}, \boldsymbol{T}^T \boldsymbol{1}_{N_m} = \boldsymbol{1}_{N_{m'}}\}$ is the set of valid permutation matrices.

Given M modalities' samples $\{\boldsymbol{X}_m \in \mathbb{R}^{N_m \times D_m}\}_{m=1}^{M}$, without loss of generality, we assume that $N_1 \geq ... \geq N_M$ and solve $M-1$ QAP problems — taking $\boldsymbol{K}_1(\boldsymbol{X}_1)$ as the reference and aligning other kernel matrices to it. Then, the problem in (1) becomes

$$\begin{aligned} &\max_{\boldsymbol{G} \in \Omega, \boldsymbol{\theta}} \operatorname{tr}(\boldsymbol{G}^T \bar{\boldsymbol{K}}(\boldsymbol{X}; \boldsymbol{\theta}) \boldsymbol{G}) \\ &s.t.\ \bar{\boldsymbol{K}}(\boldsymbol{X}; \boldsymbol{\theta}) = \sum\nolimits_{m=1}^{M} \alpha_m \boldsymbol{T}_{1,m} \boldsymbol{K}_m(\boldsymbol{X}_m; \boldsymbol{\theta}_m) \boldsymbol{T}_{1,m}^T, \end{aligned} \tag{3}$$

where $\boldsymbol{T}_{1,1} = \boldsymbol{I}_{N_1}$ and other $\boldsymbol{T}_{1,m}$'s ($m \neq 1$) are derived by solving the QAP problems.

Such an "alignment-then-clustering" strategy is challenging in practice. When the numbers of samples (the N_m's) are large, solving the QAP problems is time-consuming because of their NP-hardness. The solutions often suffer from the identifiability issue and are sensitive to the noise of data. As a result, the sub-optimal alignment may lead to catastrophic error propagation, and thus poor clustering performance. To overcome these challenges, we propose the following joint alignment-clustering framework in a bi-level optimization manner:

$$\begin{aligned} &\underbrace{\max_{\boldsymbol{G} \in \Omega, \boldsymbol{\theta}} \operatorname{tr}(\boldsymbol{G}^T \bar{\boldsymbol{K}} \boldsymbol{G})}_{\text{Fused kernel clustering}} \\ &s.t.\ \bar{\boldsymbol{K}} = \underbrace{\arg\max_{\boldsymbol{T}_m \in \Pi_m, \boldsymbol{K}} \sum\nolimits_m \alpha_m \operatorname{tr}(\boldsymbol{K}_m^T(\boldsymbol{X}_m; \boldsymbol{\theta}_m) \boldsymbol{T}_m^T \boldsymbol{K} \boldsymbol{T}_m)}_{\text{Multi-kernel alignment and fusion}}. \end{aligned} \tag{4}$$

Here, the upper-level problem corresponds to the clustering problem based on the fused kernel. The lower-level problem aligns the kernel matrices jointly, which optimizes the alignment matrices $\{\boldsymbol{T}_m\}_{m=1}^{T}$ and outputs the corresponding fused kernel $\bar{\boldsymbol{K}}$.

[1] In the following content, we may represent the kernel matrix as $\boldsymbol{K}_m$ for convenience.

The main differences between our joint alignment-clustering framework and the "alignment-then-clustering" strategy include three points: *i*) The variables of the upper-level problem, i.e., $\boldsymbol{\theta}$, are involved in the lower-level problem. As a result, we need to solve these two problems iteratively. *ii*) We do not set a reference modality, so that the fused kernel $\bar{\boldsymbol{K}}$ and the alignment matrices are learned jointly, and the size of the fused kernel can be set with high flexibility. *iii*) Denote the feasible domain of each $\boldsymbol{T}_m$ as Π_m. Instead of setting Π_m as a set of strict permutation matrices, we consider relaxing it to a set of doubly-stochastic matrices and thus avoid to solve QAP problems. In the following content, we will show that this joint alignment-clustering framework can be implemented efficiently based on the Gromov-Wasserstein distance.

3.2 Fusing kernels as calculating a weighted Gromov-Wasserstein barycenter

Gromov-Wasserstein distance is proposed in [37, 46], which is a natural extension of classic optimal transport theory [51] and provides a valid metric for metric-measure spaces (mm-spaces).

Definition 3.1. *Let $\mathcal{X}_{d_x,p_x}$ and $\mathcal{Y}_{d_y,p_y}$ be two metric measure spaces, where d_x is the metric defined in the space $\mathcal{X}$, and p_x is a probability measure defined on $\mathcal{X}$ (with $\mathcal{Y}_{d_y,p_y}$ defined in the same way). The Gromov-Wasserstein distance $D_{gw}(\mathcal{X}_{d_x,p_x}, \mathcal{Y}_{d_y,p_y})$ is defined as*

$$\begin{aligned} d_{gw}(\mathcal{X}_{d_x,p_x}, \mathcal{Y}_{d_y,p_y}) &:= \inf_{\pi\in\Pi(p_x,p_y)} \mathbb{E}_{(x,y,x',y')\sim\pi\times\pi}[r_{x,y,x',y'}] \\ &= \inf_{\pi\in\Pi(p_x,p_y)} \int_{\mathcal{X}^2\times\mathcal{Y}^2} r_{x,y,x',y'}\pi(x,y)\pi(x',y')\mathrm{d}x\mathrm{d}y\mathrm{d}x'\mathrm{d}y', \end{aligned} \tag{5}$$

where $r_{x,y,x',y'} = |d_x(x,x') - d_y(y,y')|^2$ is relational distance that measures the discrepancy between the sample pairs, and $\Pi(p_x,p_y) = \{\pi(x,y) \geq 0 | \int_{\mathcal{Y}} \pi(x,y)\mathrm{d}y = p_x, \int_{\mathcal{X}} \pi(x,y)\mathrm{d}x = p_y\}$ is the set of all probability measures on $\mathcal{X}\times\mathcal{Y}$ with p_x and p_y as marginals.

According to the above definition, the GW distance corresponds to the minimum expectation of the relational loss. The optimal joint distribution π^* corresponding to the GW distance is called the optimal transport plan (or coupling) between p_x and p_y.

Given the samples of the two mm-spaces, e.g., $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^I \subset \mathcal{X}$ and $\boldsymbol{Y} = \{\boldsymbol{y}_j\}_{j=1}^J \subset \mathcal{Y}$, whose empirical sample distributions are uniform (i.e., $\hat{\boldsymbol{p}}_x = \frac{1}{I}\mathbf{1}_I$ and $\hat{\boldsymbol{p}}_y = \frac{1}{J}\mathbf{1}_J$), the empirical Gromov-Wasserstein distance between the samples can be defined as

$$\begin{aligned} &\hat{d}_{\mathrm{gw}}(\boldsymbol{C}_X, \boldsymbol{C}_Y) \\ &:= \min_{\boldsymbol{T}\in\Pi(\hat{\boldsymbol{p}}_x,\hat{\boldsymbol{p}}_y)} \sum_{i,i'=1}^I \sum_{j,j'=1}^J |c_{ii'}^X - c_{jj'}^Y|^2 T_{ij}T_{i'j'} \\ &= \min_{\boldsymbol{T}\in\Pi(\hat{\boldsymbol{p}}_x,\hat{\boldsymbol{p}}_y)} \mathrm{tr}((\boldsymbol{C}_X \odot \boldsymbol{C}_X)\hat{\boldsymbol{p}}_x\mathbf{1}_J^T\boldsymbol{T}^T) + \\ &\qquad \mathrm{tr}(\boldsymbol{T}^T\mathbf{1}_I\hat{\boldsymbol{p}}_y^T(\boldsymbol{C}_Y \odot \boldsymbol{C}_Y)^T) - 2\mathrm{tr}(\boldsymbol{C}_Y^T\boldsymbol{T}^T\boldsymbol{C}_X\boldsymbol{T}) \\ &\Leftrightarrow \max_{\boldsymbol{T}\in\Pi(\hat{\boldsymbol{p}}_x,\hat{\boldsymbol{p}}_y)} \mathrm{tr}(\boldsymbol{C}_Y^T\boldsymbol{T}^T\boldsymbol{C}_X\boldsymbol{T}), \end{aligned} \tag{6}$$

where $\odot$ represents the Hadamard product. $\boldsymbol{C}_X = [c_{ii'}^X] \in \mathbb{R}^{I\times I}$ and $\boldsymbol{C}_Y = [c_{jj'}^Y] \in \mathbb{R}^{J\times J}$ are two relation matrices constructed by the samples, and each element $c_{ii'}^X$ indicates the relation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i'}$ quantitatively (and $c_{jj'}^Y$ works in the same way). For the samples, $\boldsymbol{C}_X$ and $\boldsymbol{C}_Y$ can be their distance matrices [35, 66], kernel matrices [36], or adjacency matrices [47, 65] (if the graph structures of the samples are available). The matrix $\boldsymbol{T}$ is restricted to be a doubly-stochastic matrix, i.e., $\boldsymbol{T} \in \Pi(\hat{\boldsymbol{p}}_x, \hat{\boldsymbol{p}}_y)$ and $\Pi(\hat{\boldsymbol{p}}_x, \hat{\boldsymbol{p}}_y) = \{\boldsymbol{T} \geq \mathbf{0} | \boldsymbol{T}\mathbf{1}_J = \hat{\boldsymbol{p}}_x, \boldsymbol{T}^T\mathbf{1}_I = \hat{\boldsymbol{p}}_y\}$. The optimal solution, denoted as $\boldsymbol{T}^*$, is called optimal transport matrix, which can be viewed as a joint distribution of the samples (i.e., $\boldsymbol{X}\times\boldsymbol{Y}$).

As shown in third row of (6), the optimization problem of the GW distance can be rewritten in a matrix format [42, 64]. Moreover, the first two terms are constant because $\boldsymbol{T}\mathbf{1}_J = \hat{\boldsymbol{p}}_x$ and $\boldsymbol{T}^T\mathbf{1}_I = \hat{\boldsymbol{p}}_y$. As a result, when computing $\hat{d}_{\mathrm{gw}}$, the objective function is the same with the QAP problem. However, the variable $\boldsymbol{T}$ is relaxed from a permutation matrix to a doubly-stochastic matrix, which simplifies the problem significantly and enriches our choice on optimization algorithms. Such a relaxation does not undermine the power of the GW distance on data alignment — the $\boldsymbol{T}^*$ indicates the joint distribution of the samples, and accordingly, its element T_{ij} represents the coherency probability of $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$. In other words, $\boldsymbol{T}^*$ achieves a "soft" assignment of the samples, matching $\boldsymbol{y}_j$ with $\boldsymbol{x}_i$ with a probability T_{ij}. The larger T_{ij} is, the more deterministic the matching result is. Due to its capability of data alignment, the GW distance has been applied in various matching tasks successfully, e.g., graph matching [65], shape matching [35], and so on.

When multiple sample sets are available, e.g., the unaligned multi-modal data in our study, we can achieve their joint alignment based on the GW distance as well. In particular, given the M kernel matrices $\{\boldsymbol{K}_m\}_{m=1}^M$, we can derive their weighted Gromov-Wasserstein barycenter [42] as follows:

$$\begin{aligned} &\bar{\boldsymbol{K}} = \arg\min_{\boldsymbol{K}} \sum_{m=1}^M \alpha_m \hat{d}_{gw}(\boldsymbol{K}, \boldsymbol{K}_m), \\ &\Leftrightarrow \arg\max_{\{\boldsymbol{T}_m\in\Pi(\bar{\boldsymbol{p}},\hat{\boldsymbol{p}}_m)\}_{m=1}^M, \boldsymbol{K}} \sum_{m=1}^M \alpha_m \big(2\mathrm{tr}(\boldsymbol{K}_m^T\boldsymbol{T}_m^T\boldsymbol{K}\boldsymbol{T}_m) \\ &\qquad\qquad - \bar{\boldsymbol{p}}^T(\boldsymbol{K}\odot\boldsymbol{K})\bar{\boldsymbol{p}}\big), \end{aligned} \tag{7}$$

where $\boldsymbol{\alpha} = [\alpha_m] \in \Delta^{M-1}$, $\odot$ represent the Hadamard product. According to the definition in (7), the matrix $\bar{\boldsymbol{K}} \in \mathbb{R}^{L\times L}$ is the weighted GW barycenter of the observed $\boldsymbol{K}_m$'s if only the sum of the weighted distances to $\boldsymbol{K}_m$'s is minimized. Note that, two hyperparameters should be predefined manually: *i*) the size of the barycenter (i.e., L); and *ii*) the empirical distribution associated with the barycenter (i.e., $\bar{\boldsymbol{p}} \in \Delta^{L-1}$), before we can calculate the GW barycenter. In the following content, we will show that L can be much smaller than the number of samples, and we can set $\bar{\boldsymbol{p}}$ to be a uniform distribution just as the way [42, 64] have done.

Replacing the GW distance with its equivalent optimization problem, we can reformulate the GW barycenter as shown in the second row of (7). It is easy to find that the weighted GW barycenter problem is coincident to the multi-kernel alignment and fusion problem in (4). For each observed kernel, the optimal transport matrix helps to align it to the barycenter. In particular, denote $\mathcal{L}$ as the objective function in (7), and suppose that the optimal transport matrices $\{\boldsymbol{T}_m^*\}_{m=1}^M$ are available. Based on the first-order optimality condition, the barycenter is derived as the weighted sum of the kernel matrices aligned by the optimal transport matrices:

$$\frac{\partial\mathcal{L}}{\partial\bar{\boldsymbol{K}}} = \mathbf{0} \quad \Rightarrow \quad \bar{\boldsymbol{K}} = \frac{1}{\bar{\boldsymbol{p}}\bar{\boldsymbol{p}}^T}\sum_{m=1}^M \alpha_m \boldsymbol{T}_m^*\boldsymbol{K}_m(\boldsymbol{T}_m^*)^T. \tag{8}$$

Therefore, the weighted GW barycenter problem is the lower-level problem of the joint alignment-clustering framework in (4).

3.3 Gromov-Wasserstein clustering

For the upper-level clustering problem in (4), we now revisit it from the viewpoint of Gromov-Wasserstein distance. In particular, according to (6), we have the following proposition:

Proposition 3.2. *For the clustering problem* $\max_{G\in\Omega} tr(G^T\bar{K}G)$, *if the indicator* G *is a doubly-stochastic matrix, i.e.,* $\Omega = \Pi(\bar{p}, \frac{1}{d}\mathbf{1}_d) = \{G \geq \mathbf{0} | G\mathbf{1}_d = \bar{p}, G^T\mathbf{1}_L = \frac{1}{d}\mathbf{1}_d\}$, *then we can get that*

$$\max_{G\in\Omega} tr(G^T\bar{K}G) \Leftrightarrow \max_{G\in\Omega} tr(G^T\bar{K}GI_d) \Leftrightarrow \hat{d}_{gw}(\bar{K}, I_d). \quad (9)$$

This equivalence is firstly applied in [65], which helps to achieve encouraging performance on clustering tasks like graph partitioning. Proposition 3.2 indicates that we can solve the clustering problem driven by the modularity maximization principle by means of the GW distance. Recently, the work in [6] demonstrates in theory that computing $\hat{d}_{\mathrm{gw}}(\boldsymbol{K}, \boldsymbol{I}_d)$ can be regarded as a solution of the generalized spectral clustering given the kernel matrix $\boldsymbol{K}$. Especially, when the kernel is a heat kernel and the clusters have comparable size, computing $\hat{d}_{\mathrm{gw}}(\boldsymbol{K}, \boldsymbol{I}_d)$ with $d = 2$ has achieved the well-known Fiedler partitioning. The clustering problem can be further simplified with higher flexibility and efficiency when the kernel is the GW barycenter of multiple kernels.

3.3.1 ***Reduce the problem size for efficiency.*** Plugging the $\bar{K}$ in (8) into the upper-level clustering problem in (4), we have

$$\begin{aligned}
&\underbrace{\max_{G\in\Pi(\bar{p},\frac{1}{d}\mathbf{1}_d)} \mathrm{tr}(G^T\bar{K}G)}_{\text{Multi-modal clustering}} \\
&= \max_{G\in\Pi(\bar{p},\frac{1}{d}\mathbf{1}_d)} \frac{1}{\bar{p}\bar{p}^T} \sum_{m=1}^{M} \alpha_m \mathrm{tr}(G^T T_m^* K_m \underbrace{(T_m^*)^T G}_{G_m}) \\
&\leq \underbrace{\max_{\{G_m\in\Pi(p_m,\frac{1}{d}\mathbf{1}_d)\}_{m=1}^M} \frac{1}{\bar{p}\bar{p}^T} \sum_{m=1}^{M} \alpha_m \mathrm{tr}(G_m^T K_m G_m)}_{\text{Clustering each modality independently}}.
\end{aligned} \quad (10)$$

Here, clustering each modality independently means learning modality-specific indicators, denoted as $\{G_m\}_{m=1}^M$, without considering whether these indicators own the same clustering structure or not. The clustering achieved by the fused kernel actually imposes a common clustering structure on these indicators, i.e., $G_m = (T_m^*)^T G$ for $m = 1, ..., M$.[2] In other words, each modality-specific indicator is factorized into two components: the modality-specific alignment matrix T_m^* and a shared clustering indicator G. Imposing this structural constraint makes the multi-modal clustering works as a lower bound of independent single modality clustering.

This factorization model $G_m = (T_m^*)^T G$ provides important evidence for determining the size of the GW barycenter $\bar{K}$. The rank of G_m is at most d (i.e., the number of clusters), so it is sufficient to set the size of the GW barycenter L to be d, which avoids introducing redundant structural information. As a result, the upper-level clustering problem in (4) can be transformed into a small-scale GW distance problem, where both G and $\bar{K}$ are with size $(d \times d)$.

[2]Note that, each modality-specific indicator $G_m = (T_m^*)^T G$ because both T_m^* and G are doubly-stochastic matrices, each modality-specific indicator $G_m = (T_m^*)^T G$ is a doubly-stochastic matrix as well, whose feasible domain is $\Pi(p_m, \frac{1}{d}\mathbf{1}_d)$.

3.3.2 ***Design the clustering loss with high flexibility.*** As shown in Proposition 3.2, the clustering problem is equivalent to computing $\hat{d}_{\mathrm{gw}}(\bar{K}, I_d)$. Furthermore, when the $\bar{K}$ is a GW barycenter, other loss functions are also applicable because both the GW distance and the GW barycenter own a useful property — permutation-invariance [64]. In particular, given a matrix $\boldsymbol{K}$ and its permutation $\boldsymbol{P}\bar{\boldsymbol{K}}\boldsymbol{P}^T$, its GW distance to an arbitrary matrix $\boldsymbol{K}'$ satisfies:

$$\hat{d}_{\mathrm{gw}}(\boldsymbol{K}, \boldsymbol{K}') = \hat{d}_{\mathrm{gw}}(\boldsymbol{P}\bar{\boldsymbol{K}}\boldsymbol{P}^T, \boldsymbol{K}'), \quad (11)$$

where $\boldsymbol{P} \in \mathcal{P}$ is a random permutation matrix. For the GW barycenter problem (7), if $\bar{\boldsymbol{K}}$ is its optimal solution, then its permutation $\boldsymbol{P}\bar{\boldsymbol{K}}\boldsymbol{P}^T$ will become another optimal solution. As a result, due to the permutation-invariant merit of the GW barycenter $\bar{\boldsymbol{K}}$, the permutation-invariance in the clustering loss becomes nonobligatory. Thus, besides $\hat{d}_{\mathrm{gw}}(\bar{\boldsymbol{K}}, \boldsymbol{I}_d)$, other loss functions like mean-square error (MSE) and cross-entropy loss are applicable.

In summary, plugging the GW barycenter problem in (7) into the joint alignment-clustering framework in (4), we obtain the proposed Gromov-Wasserstein multi-modal alignment and clustering (GWMAC) model as follows:

$$\begin{aligned}
&\min_{\boldsymbol{\theta}} \mathrm{loss}(\bar{\boldsymbol{K}}(\boldsymbol{\theta}), \boldsymbol{I}_d)) - \gamma H(\boldsymbol{\alpha}) \\
&s.t.\ \bar{\boldsymbol{K}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{K}} \sum_{m=1}^{M} \alpha_m \hat{d}_{\mathrm{gw}}(\boldsymbol{K}, \boldsymbol{K}_m(\boldsymbol{X}_m; \boldsymbol{\theta}_m)),
\end{aligned} \quad (12)$$

where the size of the barycenter $\bar{\boldsymbol{K}}$ is $d \times d$, and $\bar{\boldsymbol{p}} = \frac{1}{d}\mathbf{1}_d$. For the loss function $\mathrm{loss}(\bar{\boldsymbol{K}}(\boldsymbol{\theta}), \boldsymbol{I}_d))$, we consider three options:

- **The GW-based loss:** $\min_{G\in\Pi(\bar{p},\frac{1}{d}\mathbf{1}_d)} -\mathrm{tr}(G^T\bar{\boldsymbol{K}}(\boldsymbol{\theta})G)$.
- **The Mean-Squared Error (MSE):** $\|\bar{\boldsymbol{K}}(\boldsymbol{\theta}) - \boldsymbol{I}_d\|_F^2$.
- **The cross-entropy (CE) loss:** When $\mathbf{0} \leq \bar{\boldsymbol{K}}(\boldsymbol{\theta}) \leq \mathbf{1}$, we can apply $\mathrm{tr}(\log \bar{\boldsymbol{K}}(\boldsymbol{\theta})) + \mathrm{tr}((1 - \boldsymbol{I}_d)\log(1 - \bar{\boldsymbol{K}}(\boldsymbol{\theta})))$.

Additionally, we introduce $H(\boldsymbol{\alpha}) = -\langle \boldsymbol{\alpha}, \log \boldsymbol{\alpha} \rangle$ as the entropy of the modalities' weights, which would avoid learning the clustering structure just from a single modality. The significance of this term is controlled by the hyperparameter γ and $\gamma \geq 0$.

For each kernel matrix $\boldsymbol{K}_m(\boldsymbol{X}_m; \boldsymbol{\theta}_m)$, we construct it based on the latent codes obtained by the encoder. Given arbitrary two samples, i.e., $\boldsymbol{x}_i^m, \boldsymbol{x}_j^m \in \boldsymbol{X}_m$, the corresponding element of $\boldsymbol{K}_m(\boldsymbol{X}_m; \boldsymbol{\theta}_m)$, i.e., K_{ij}^m, is modeled as

$$K_{ij}^m = K_\sigma(f_{\boldsymbol{\theta}_m}(\boldsymbol{x}_i^m), f_{\boldsymbol{\theta}_m}(\boldsymbol{x}_j^m)),\ m = 1, ..., M. \quad (13)$$

where $f_{\boldsymbol{\theta}_m}$ represents the encoder of the m-th modality and parameterized by $\boldsymbol{\theta}_m$, and $K_\sigma(\cdot, \cdot)$ is a predefined kernel function whose bandwidth is σ. In this work, we implement each encoder as a multi-layer perceptron (MLP) model and the kernel function as the radial basis function (RGB) kernel.

4 LEARNING ALGORITHM

4.1 Alternating optimization

To achieve a trade-off between performance and efficiency, we propose an alternating optimization strategy for (12), first calculating the GW barycenter and the associated optimal transport matrices, and then updating the model parameters via stochastic gradient descent (SGD). Specifically, our algorithm involves the following two steps at the t-th iteration.

Algorithm 1 Conditional gradient algorithm for $\hat{d}_{\text{gw}}(\bar{\boldsymbol{K}}, \boldsymbol{K})$

Input: $\bar{\boldsymbol{K}} \in \mathbb{R}^{d\times d}$, $\boldsymbol{K} \in \mathbb{R}^{N\times N}$, $\bar{\boldsymbol{p}} \in \Delta^{d-1}$, and $\boldsymbol{p} \in \Delta^{N-1}$.
Initialize $\boldsymbol{T} = \bar{\boldsymbol{p}}\boldsymbol{p}^T$.
while not converge **do**
  (*i*) **Apply the network flow algorithm:**
  $\widetilde{\boldsymbol{T}} = \arg\max_{\boldsymbol{T}\in\Pi(\bar{\boldsymbol{p}},\boldsymbol{p})} \text{tr}(\boldsymbol{K}^T\boldsymbol{T}^T\bar{\boldsymbol{K}}\boldsymbol{T})$.
  (*ii*) **Apply the line search method:**
  $a = -2\text{tr}(\boldsymbol{K}^T\widetilde{\boldsymbol{T}}^T\bar{\boldsymbol{K}}\widetilde{\boldsymbol{T}})$, $b = \text{tr}((\bar{\boldsymbol{K}}\odot\bar{\boldsymbol{K}})\bar{\boldsymbol{p}}\boldsymbol{p}^T + \bar{\boldsymbol{p}}\boldsymbol{p}^T(\boldsymbol{K}\odot\boldsymbol{K})^T)$
  $c = -2(\text{tr}(\boldsymbol{K}^T\boldsymbol{T}^T\bar{\boldsymbol{K}}\widetilde{\boldsymbol{T}}) + \text{tr}(\boldsymbol{K}^T\widetilde{\boldsymbol{T}}^T\bar{\boldsymbol{K}}\boldsymbol{T}))$.
  if $a > 0$ **then**
    $\tau = \min(1, \max(0, \frac{-(b+c)}{2a}))$
  else
    $\tau = 1$ if $a + b + c < 0$ else $\tau = 0$
  end if
  (*iii*) **Update OT matrix:** $\boldsymbol{T} \leftarrow (1-\tau)\boldsymbol{T} + \tau\widetilde{\boldsymbol{T}}$
end while
Output: $\boldsymbol{T}^* := \boldsymbol{T}$.

4.1.1 ***Update GW barycenters.*** Given current model parameters, we first solve the lower-level problem in (12) to calculate the GW barycenter based on the kernel matrices.

$$\min_{\boldsymbol{K}} \sum_{m=1}^{M} \alpha_m^{(t)} \hat{d}_{\text{gw}}(\boldsymbol{K}, \boldsymbol{K}_m(\boldsymbol{X}_m; \boldsymbol{\theta}_m^{(t)})). \tag{14}$$

Solving this problem involves an inner iteration with L steps. Given the current barycenter $\bar{\boldsymbol{K}}^{(\ell)}$ at the ℓ-th inner step, we first compute GW distances M times, i.e., $\hat{d}_{\text{gw}}(\bar{\boldsymbol{K}}^{(\ell)}, \boldsymbol{K}_m(\boldsymbol{X}_m; \boldsymbol{\theta}_m^{(t)})$, to obtain the optimal transport matrices $\{\boldsymbol{T}_m^*\}_{m=1}^M$. Then, we update the barycenter via (8) with the optimal transport matrices.

Although the GW distance corresponds to a nonconvex non-smooth optimization problem, many algorithms can be applied to solve it efficiently, e.g., the proximal gradient algorithm [42, 67], the Bregman ADMM algorithm [52, 64], and so on. In this work, we apply the conditional gradient algorithm proposed in [47] to pursue sparse optimal transport matrices. The specific scheme to calculate the GW distance is summarized in Algorithm 1.

4.1.2 ***Update model parameters.*** Plugging the $\bar{\boldsymbol{K}}$ calculated in the first step into the upper-level problem in (12), we have

$$\min_{\boldsymbol{\theta}_m, \boldsymbol{\alpha}\in\Delta^{M-1}} \text{loss}\Big(\frac{1}{\bar{\boldsymbol{p}}\bar{\boldsymbol{p}}^T}\sum_{m=1}^{M}\alpha_m\boldsymbol{T}_m^*\boldsymbol{K}_m(\boldsymbol{\theta}_m)(\boldsymbol{T}_m^*)^T, \boldsymbol{I}_d\Big) - \gamma H(\boldsymbol{\alpha}). \tag{15}$$

We can update $\{\boldsymbol{\theta}_m\}_{m=1}^M$ efficiently by the SGD algorithm, where the gradient is calculated via the backpropagation. Meanwhile, we need to ensure that the updated values of $\boldsymbol{\alpha}$ are in $(M-1)$-Simplex. To achieve this, we project the $\tilde{\boldsymbol{\alpha}}$ obtained by the SGD back to the $(M-1)$-Simplex by optimizing $\min_{\boldsymbol{\alpha}\in\Delta^{M-1}} \|\boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}}\|_2^2$.

Details of our GWMAC method are shown in Algorithm 2. The output of GWMAC method involves three parts: (*i*) The optimal transport matrices $\{\boldsymbol{T}_m \in \Pi(\frac{1}{d}\boldsymbol{1}_d, \frac{1}{N_m}\boldsymbol{1}_{N_m})\}_{m=1}^M$ that provide us with the joint distribution of the samplers per modality and the clustering result we are interested in. Specifically, the n-th sample of the m-th modality is in the i^*-th cluster if $i^* = \arg\max_{i\in\{1,..,d\}} T_{ni,m}^*$ given $\boldsymbol{T}_m^* = [T_{ni,m}^*]$. (*ii*) The vector $\boldsymbol{\alpha}$ that indicates the significance of different modalities, helps us to find the useful modalities for our clustering task, and at the same time, suppresses the negative influences of the useless modalities. (*iii*) The encoders $f_{\boldsymbol{\theta}_m}$ that make our model inductive. As a result, we can leverage the encoders to represent new-coming data and achieve other downstream applications like multi-modal classification, besides clustering the observed multi-modal data.

Algorithm 2 Algorithm for GWMAC

Input: Multi-modal data and their sample distributions $\{\boldsymbol{X}_m \in \mathbb{R}^{N_m\times D_m}, \boldsymbol{p}_m = \frac{1}{N_m}\boldsymbol{1}_{N_m}\}_{m=1}^M$, the predefined number of clusters d, and $\bar{\boldsymbol{p}} = \frac{1}{d}\boldsymbol{1}_d$.
while not converged **do**
  Compute $\{\boldsymbol{K}_m(\boldsymbol{B}_m; \boldsymbol{\theta}_m)\}_{m=1}^M$ for a batch $\{\boldsymbol{B}_m \subset \boldsymbol{X}_m\}_{m=1}^M$.
  (*i*) **Solve the GW barycenter problem:**
  Initialize $\bar{\boldsymbol{K}} = \boldsymbol{I}_d$
  for $\ell = 1, ..., L$ **do**
    for $m = 1, ..., M$ **do**
      Compute $\hat{d}_{\text{gw}}(\bar{\boldsymbol{K}}, \boldsymbol{K}_m)$ via Algorithm 1 and obtain $\boldsymbol{T}_m^*$.
    end for
    Update $\bar{\boldsymbol{K}}$ by (8).
  end for
  (*ii*) **Update model parameters:**
  Solve (15) via SGD and obtain $\{\boldsymbol{\theta}_m\}_{m=1}^M$ and $\tilde{\boldsymbol{\alpha}}$.
  Update $\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}}\|_2^2$.
end while
Output: $\{\boldsymbol{T}_m^*\}_{m=1}^M$, $\boldsymbol{\alpha}$, and the encoders $\{f_{\boldsymbol{\theta}_m}\}_{m=1}^M$.

Table 1: Summary of the multi-modal datasets

Datasets	Size	Views	Dimensions	Class
HandWritten	2000	6	[76, 216, 64, 240, 47, 6]	10
Caltech 7	1474	6	[48, 40, 254, 1984, 512, 928]	7
Movies	617	2	[1878, 1398]	17
ORL	400	2	[288, 288]	40
Prokaryotic	551	3	[438, 3, 393]	4

Denote the number of samples per batch as B, and the expected number of the clusters as d, then the computational complexity of our GWMAC method is $O(LMB^2d)$, which is mainly contributed by computing the GW distances M times in L inner GW barycenter iterations. Fortunately, we often have $L, M, d, B \ll N$ in practice where N is the total number of samples. Moreover, the computation of the GW distance can be accelerated by various methods, e.g., applying the "divide-and-conquer" strategy in [65], or imposing low-rank structures to the kernel matrices and the optimal transport matrices [45], and so on. As a result, the computational complexity of our GWMAC method can be further reduced to $O(LMBd\log B)$. To our knowledge, this complexity is at least comparable to that of most existing methods and only slightly higher than that of [55].

5 EXPERIMENTS

To demonstrate the feasibility and the effectiveness of our GWMAC method, we test and analyze it on several representative multi-modal datasets and compare it with state-of-the-art multi-modal clustering methods on various learning scenarios.

Table 2: The performance of different clustering methods. Here, "-" means that a method fails to obtain results in 10 hours.

Data type	Datasets	HandWritten		Caltech 7		ORL		Movies		Prokaryotic	
	Algorithms	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Well-aligned ($\beta = 0$)	MCCA	0.8269	0.7775	**0.5313**	0.4716	0.3475	0.4992	0.0989	0.0722	0.5620	0.1204
	DCCAE	0.6537	0.6216	0.4110	0.3850	0.5625	0.7373	0.1572	0.1194	0.5070	0.1827
	AttnAE	0.7505	0.6912	0.4600	0.4575	0.4600	0.6603	0.1880	0.1918	0.5390	0.2625
	MVKSC	0.6749	0.6376	0.5196	0.2537	0.3013	0.5291	**0.2285**	**0.2098**	**0.6188**	**0.3191**
	MultiNMF	**0.8882**	**0.8279**	0.4525	**0.5120**	**0.6900**	**0.8100**	0.1726	0.1856	0.5771	0.2495
50% unaligned ($\beta = 0.5$)	CPM-GAN	0.7250	0.6069	0.3472	0.3151	0.1987	0.3703	0.1210	0.1753	0.3793	**0.3294**
	MVC-UM	-	-	**0.3958**	0.3838	**0.5863**	**0.7586**	0.1831	0.1950	0.3950	0.0807
	GWMAC	**0.8469**	**0.8156**	0.3541	**0.5010**	0.5322	0.7068	**0.1993**	**0.2195**	**0.5515**	0.3286
100% unaligned ($\beta = 1$)	MVC-UM	-	-	0.3112	0.2456	**0.5431**	**0.7452**	0.1841	0.1953	0.4451	0.0554
	GWMAC	**0.8144**	**0.7546**	**0.3568**	**0.4945**	0.5118	0.7026	**0.1928**	**0.2138**	**0.5479**	**0.3259**

5.1 Implementation details

5.1.1 ***Datasets.*** In the following experiments, we consider five commonly-used multi-modal datasets, including HandWritten, Caltech 7, Movies, ORL, and Prokaryotic. The links of the datasets can be found at our codebase. Each of them contains well-aligned multi-modal samples and each sample owns a class label. The basic statistics of these multi-modal datasets are summarized in Table 1. To demonstrate the power of our method, we construct the unaligned multi-modal data in a controllable way. We set an unalignment ratio β in $[0, 1]$, and we randomly permute the top $\beta \times 100\%$ percentage samples of each modality in each dataset. Obviously, $\beta = 0$ corresponds to the original well-aligned data, while $\beta = 1$ leads to a totally-unaligned multi-modal dataset. Given such datasets, we set the batch size to be 400 when applying the SGD learning algorithm.

5.1.2 ***Baselines and evaluation measurements.*** For each dataset, we apply various multi-modal clustering methods, which can be categorized into two classes:

1) Five classic multi-modal clustering methods: The **MCCA** in [9] fuses the samples in all modalities into one mutual space, and concatenates the latent codes accordingly. The Deep Canonical Correlation Auto-Encoders (**DCCAE**) in [56] learns a auto-encoding network with a CCA-based objective. The **AttnAE** applies a M-head self-attention layer to obtain the latent codes of the multi-modal data, in which each head encodes one modality and the outputs of all the heads are further fused by a self-attention layer. The latent codes are learned to reconstruct the data (through M decoders). For these three methods, we apply K-means to the learned latent codes. The **MultiNMF** in [30] is a nonnegative matrix factorization (NMF) method, which learns the coefficient matrices from all modalities and regularizes them towards a shared latent clustering structure. The state-of-the-art kernel fusion clustering strategy, i.e., the multi-view kernel spectral clustering (**MVKSC**) method in [21]. All above multi-modal clustering methods are dependent on the well-aligned multi-modal data.

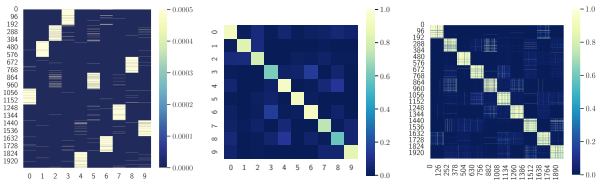
2) Two state-of-the-art methods for unaligned multi-modal data: The **CPM-GAN** in [72] is a deep learning method for partially-unaligned multi-modal data modeling, which leverages a generative adversarial network to generate the unobserved modalities of each unaligned sample conditioned on the observed modalities. The **MVC-UM** in [70] is the state-of-the-art multi-modal clustering methods applicable for both partially-unaligned and totally-unaligned multi-modal data by jointly learning the factorization models of all the modalities and the correspondence between arbitrary two modalities. Note that, the PVC in [22] is designed for clustering with two modalities and cannot be easily extended to multi-modal clustering, so we do not choose it as our baseline.

For our GWMAC method and the above baselines, we apply the grid search method to find their optimal hyperparameters, e.g., the number of epochs, the batch size, the learning rate, and so on. Additionally, we further study the influences of some key hyperparameters on our GWMAC methods in the following experiments, including the number of inner iterations L for computing GW barycenters, the bandwidth σ of kernel function, the weight γ of the entropic regularizer, and the choice of various loss functions. Following the work in [30, 70, 76], we evaluate the above clustering methods based on their clustering accuracy (ACC) and the normalized mutual information (NMI) on the datasets. For each clustering method, we run it in five trials under its optimal hyperparameter setting but with different random seeds. We report the averaged performance of each method as the final result.[3]

5.2 Comparisons and analysis

5.2.1 ***Clustering performance.*** We consider the performance of the multi-modal clustering in three data scenarios: (*i*) well-aligned multi-modal data are available; (*ii*) 50% data are unaligned; (*iii*) the data are totally-unaligned, and we test various methods in their own applicable scenarios. Table 2 lists the clustering results of all methods on the five datasets. In each data scenario, we bold the best results and underline the second best results, respectively. Our GWMAC method outperforms its main competitors (i.e., MVC-UM and CPM-GAN) in most situations, and its clustering performance is even superior or comparable to some baselines trained on the well-aligned data.

[3]For each method, we find that the standard deviation of its clustering ACC in five trials is less than 0.1. In the following tables, we can find that the gaps between the averaged performance of different methods are statistically-significant compared to the standard deviation.



(a) T_2^* (b) $\bar{K}$ (c) $\sum_m K_m$

Figure 2: Visualizations of the learning results achieved on the HandWritten dataset. Here, we sort the samples in advance for good visual effects.

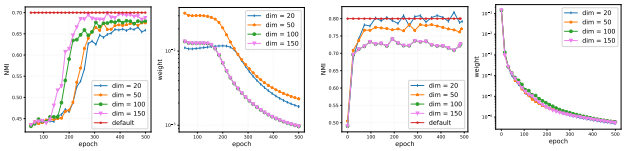
A potential reason for this phenomenon is that the optimal transport matrices align the samples across different modalities in a probabilistic way, which achieves data augmentation to some degrees. In contrast, the classic methods can just leverage the deterministic N paired samples, i.e., $\{\boldsymbol{x}_i^m, \boldsymbol{x}_i^{m'}\}_{i=1}^N$, given two sample sets $\boldsymbol{X}_m = \{\boldsymbol{x}_i^m\}_{i=1}^N$ and $\boldsymbol{X}_{m'} = \{\boldsymbol{x}_i^{m'}\}_{i=1}^N$. Meanwhile, our GWMAC learns $\boldsymbol{T}_m^*$ and $\boldsymbol{T}_{m'}^*$ to align $\boldsymbol{B}_m$ and $\boldsymbol{B}_{m'}$ to the clusters without the correspondence information, given two batches $\boldsymbol{B}_m \subset \boldsymbol{X}_m$ and $\boldsymbol{B}_{m'} \subset \boldsymbol{X}_{m'}$. Accordingly, the correspondence between $\boldsymbol{B}_m$ and $\boldsymbol{B}_{m'}$ is estimated by $(\boldsymbol{T}_m^*)^T \boldsymbol{T}_{m'}^*$. As a result, the unchanged correspondence is replaced by the probabilistic correspondence that changes during training to achieve the augmentation of the fused kernels.

For partially-unaligned and totally-unaligned multi-modal data, our GWMAC often works better than CPM-GAN and MVC-UM, especially when the datasets are complex (i.e., HandWritten, Caltech 7, and Prokaryotic). On one hand, CPM-GAN requires to learn a generative model to estimate the missed modalities of those unaligned data, which often suffers from the over-fitting issue and has a high risk of model misspecification. On the other hand, MVC-UM learns pairwise correspondence between arbitrary two modalities, which can not guarantee the consistency among more than two modalities (i.e., $\boldsymbol{x}_i^1$ matches with $\boldsymbol{x}_j^2$, $\boldsymbol{x}_j^2$ matches with $\boldsymbol{x}_k^3$, but $\boldsymbol{x}_i^1$ may not match with $\boldsymbol{x}_k^3$). Our GWMAC, however, aligns multiple modalities jointly with a single barycenter, which naturally owns better alignment consistency. Additionally, MVC-UM has high-complexity so that its efficiency may be questionable when the number of samples is large, as shown in Table 2.

Focused on the HandWritten dataset, we further visualize the learning results obtained by our method in Figure 2. In particular, Figure 2(a) shows the learned optimal transport matrix corresponding to a representative modality (i.e., $\boldsymbol{T}_2^*$). We can find that the 2,000 samples of the dataset are assigned to 10 clusters with high accuracy, which demonstrates the rationality of the optimal transport we have learned. Based on such optimal transport matrices, we can derive the fused kernel matrix and reveal the clustering structure of the data, as shown in Figure 2(b). Furthermore, we derive the latent codes of each modality through the encoders and compute the kernel matrix of the samples directly, i.e., $\{\boldsymbol{K}_m\}_{m=1}^M$. Additionally, given well-aligned multi-modal data, we visualize $\sum_{m=1}^M \boldsymbol{K}_m$ in Figure 2(c). The result reflects the clustering structure of the data with high accuracy, which demonstrates the rationality of the encoders learned by our GWMAC method.

Table 3: The performance of various methods on clustering unseen but well-aligned multi-modal data.

Algorithms / Datasets	MVC-UM ACC	MVC-UM NMI	GWMAC ACC	GWMAC NMI
HandWritten	-	-	**0.8368**	**0.8271**
Caltech 7	0.2910	0.1917	**0.3738**	**0.4845**
ORL	0.2118	0.7603	**0.5624**	**0.8284**
Movies	0.2014	0.2379	**0.2032**	**0.2613**
Prokaryotic	0.4099	0.0981	**0.6179**	**0.4496**



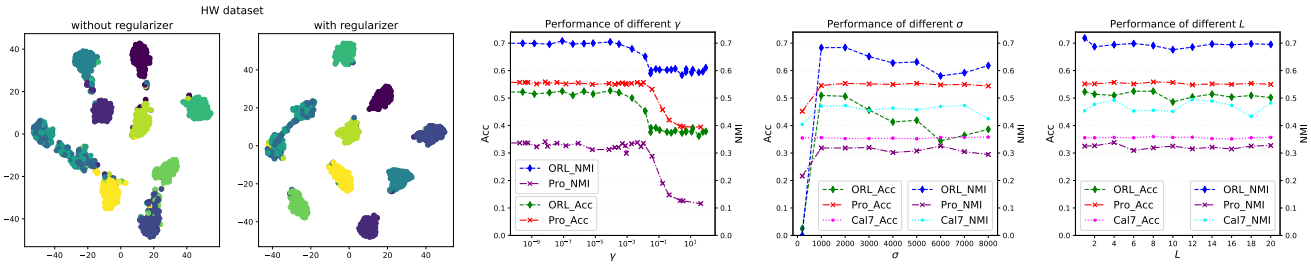
(a) ORL (b) HandWritten

Figure 3: In each subfigure, we visualize the NMI (left) and the significance of noisy modality (right) that change to the number of training epochs. In the NMI plots, the red lines are the results achieved based on the "default" data.

5.2.2 ***Inductive inference and generalization power.*** Classic methods, like MVKSC, MultiNMF, and MVC-UM, are normally dependent on transductive inference when new data comes, while our GWMAC method can achieve inductive inference by directly obtaining the latent codes of new data through the learned encoders. Therefore, our method is more efficient than the baselines in the testing phase. To demonstrate the generalization power of the encoders learned by our method, we design the following experiment. For each dataset, we split the multi-modality data randomly into 80% unaligned training data and 20% well-aligned testing data. Then we apply MVC-UM and GWMAC to learn the clustering models on the training data. For MVC-UM, we obtain the latent codes of testing data in a transductive way, i.e., fixing the learnt latent factors and optimizing the latent codes. For our GWMAC method, we obtain the latent codes of the testing data through the learnt encoders directly, and then we concatenate the latent codes of all the modalities and apply K-means to cluster the testing data. The clustering results are shown in Table 3. We can find that our GWMAC outperforms the MVC-UM significantly and consistently. This result indicates that the encoders derived by our method own good generalization power, which can obtain high-quality latent codes for new data.

5.2.3 ***Robustness to noisy modality.*** Additionally, our GWMAC method is robust to noisy modality because it learns the significance of different modalities automatically. As a result, it tends to suppress the influence of noisy modalities by reducing its significance.

To verify its robustness, we design the following experiment on ORL and HandWritten datasets. We further add a noisy modality whose samples are totally random noises to each dataset, and the number of samples in this noisy modality is just the same with



(a) Comparison on the latent codes of HandWritten data (b) The influence of γ (c) The influence of σ (d) The influence of L

Figure 4: (a) The t-SNE plots of the latent codes learned with and without the entropic regularizer. (b-d) The influences of various hyperparameters on the performance of our method for some representative datasets.

that of the original dataset while the sample dimension is set from $\{20, 50, 100, 150\}$. Applying our method to the data with the noisy modality, we visualize its performance on NMI and the significance of the noisy modality in Figure 3. We can find that as the values of the training epochs increase, the significance of the noisy modality reduces rapidly, and the performance of our method converges accordingly and approximately to the NMI achieved on the original dataset. Under different sample dimensions, we can still get the similar phenomenons even if the dimension of the noise is high. In other words, our GWMAC method is able to eliminate the effect of noisy modalities during training.

5.2.4 ***The necessity of entropic regularizer.*** Although learning the significance of modalities without any regularizer helps to remove noisy modalities as shown in Figure 3, the significance tends to be over-sparse and our GWMAC may focus too much on a single dominated modality. Therefore, we impose the entropic regularizer on the significance of modalities when learning our model. To demonstrate the necessity of the entropic regularizer, we train our encoders with and without the regularizer, respectively, on HandWritten's (well-aligned) samples. For each model, we concatenate the latent codes of different modalities and visualize the t-SNE plot of the latent codes in Figure 4(a). We can find that with the help of the regularizer, the clustering structure of the representations is less noisy. Based on the results in Figures 3 and 4(a), we need to achieve a trade-off between the robustness to noise and the usage of multi-modal information. Empirically, using small γ, i.e., $\gamma < 10^{-3}$, could lead to the stable performance as shown in Figure 4(b).

5.2.5 ***Robustness to other hyperparameters.*** Besides γ, we further consider three more key hyperparameters and quantitatively analyze their influences on our method, including the bandwidth σ of the kernel function, the inner iteration number L for computing barycenters, and the types of loss function. For each hyperparameter, we fix the remaining hyperprameters and test our method under different settings. Figures 4(c) and 4(d) visualize the performance of our method under different σ's and L's, respectively. We can find that our GWMAC method is robust to the changes of σ and L, whose performance is relatively stable when the hyperparameters change in wide ranges. With respect to the loss function in (12), we consider three kinds of loss function: GW distance, MSE loss, and CE loss. Table 4 shows the quantitative performance of our method using different loss functions, and we can find that the learning results obtained based on the GW distance and the MSE loss are comparable, while using the CE loss may lead to the degradation of the performance.

Table 4: The performance under different loss functions

Datasets	Caltech 7		ORL		Prokaryotic	
Loss	ACC	NMI	ACC	NMI	ACC	NMI
$\hat{d}_{\text{gw}}$	**0.3596**	0.4826	0.5348	0.7082	**0.5587**	**0.3409**
MSE	**0.3596**	**0.4915**	**0.5496**	**0.7205**	**0.5587**	**0.3409**
CE	0.3507	0.4697	0.4503	0.6608	0.5534	0.3062

6 CONCLUSION

We have developed a novel Gromov-Wasserstein multi-modal alignment and clustering (GWMAC) method. This method achieves the alignment and the clustering of multi-modal data jointly, and can be applied to both aligned and unaligned multi-modal data. The proposed method outperforms state-of-the-art multi-modal clustering methods in various datasets. In the future, we will further consider adding a feature selection mechanism before clustering to eliminate abnormal or noisy samples, which may leads to a barycenter paradigm under the partial Gromov-Wasserstein distance. We also plan to extend our Gromov-Wasserstain barycenter to a fused Gromov-Wasserstain barycenter, leveraging not only the structural information of data pairs but also the attribute information of data points when aligning different modalities.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 62106271), Beijing Outstanding Young Scientist Program (No. BJJWZYJH012019100020098), and the Research Funds of Renmin University of China (the Fundamental Research Funds for the Central Universities). We also would like to thank the supports from the Beijing Key Laboratory of Big Data Management and Analysis Methods, the Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, and the Public Policy and Decision-making Research Lab of RUC.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.

[2] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. 2015. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing* 37, 2 (2015), A1111–A1138.

[3] Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, and Raman Arora. 2019. Deep Generalized Canonical Correlation Analysis. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. 1–6.

[4] Chandan Biswas, Debasis Ganguly, Dwaipayan Roy, and Ujjwal Bhattacharya. 2019. Privacy Preserving Approximate K-Means Clustering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1321–1330. https://doi.org/10.1145/3357384.3357969

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

[6] Samir Chowdhury and Tom Needham. 2021. Generalized spectral clustering via Gromov-Wasserstein learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 712–720.

[7] C Mario Christoudias, Raquel Urtasun, and Trevor Darrell. 2008. Multi-view learning in the presence of view disagreement. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*. 88–96.

[8] Marco Cuturi and Arnaud Doucet. 2014. Fast computation of Wasserstein barycenters. In *International conference on machine learning*. PMLR, 685–693.

[9] Alain de Cheveigné, Giovanni M. Di Liberto, Dorothée Arzounian, Daniel D.E. Wong, Jens Hjortkjær, Søren Fuglsang, and Lucas C. Parra. 2019. Multiway canonical correlation analysis of brain data. *NeuroImage* 186 (2019), 728–740. https://doi.org/10.1016/j.neuroimage.2018.11.026

[10] Virginia R De Sa, Patrick W Gallagher, Joshua M Lewis, and Vicente L Malave. 2010. Multi-view kernel construction. *Machine learning* 79, 1-2 (2010), 47–71.

[11] Zhengming Ding and Yun Fu. 2014. Low-rank common subspace for multi-view learning. In *2014 IEEE international conference on Data Mining*. IEEE, 110–119.

[12] Liang Du, Peng Zhou, Lei Shi, Hanmo Wang, Mingyu Fan, Wenjian Wang, and Yi-Dong Shen. 2015. Robust multiple kernel K-means using $\ell 2$; 1-norm. In *Proceedings of the 24th International Conference on Artificial Intelligence*. 3476–3482.

[13] Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. 2021. ANIMC: A Soft Approach for Auto-weighted Noisy and Incomplete Multi-view Clustering. *IEEE Transactions on Artificial Intelligence* (2021).

[14] Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. 2021. Unbalanced Incomplete Multi-view Clustering via the Scheme of View Evolution: Weak Views are Meat; Strong Views do Eat. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2021).

[15] Maurizio Filippone, Francesco Camastra, Francesco Masulli, and Stefano Rovetta. 2008. A survey of kernel and spectral methods for clustering. *Pattern recognition* 41, 1 (2008), 176–190.

[16] Quanxue Gao, Wei Xia, Zhizhen Wan, Deyan Xie, and Pu Zhang. 2020. Tensor-SVD based graph learning for multi-view subspace clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 3930–3937.

[17] Mehmet Gönen and Adam A Margolin. 2014. Localized data fusion for kernel k-means clustering with application to cancer biology. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*. 1305–1313.

[18] Dongyan Guo, Jian Zhang, Xinwang Liu, Ying Cui, and Chunxia Zhao. 2014. Multiple kernel learning based multi-view spectral clustering. In *2014 22nd International conference on pattern recognition*. IEEE, 3774–3779.

[19] Yuhong Guo and Min Xiao. 2012. Cross language text classification via subspace co-regularized multi-view learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*. 915–922.

[20] Jingrui He and Rick Lawrence. 2011. A graph-based framework for multi-task multi-view learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*. 25–32.

[21] Lynn Houthuys, Rocco Langone, and Johan A.K. Suykens. 2018. Multi-View Kernel Spectral Clustering. *Information Fusion* 44 (2018), 46–56. https://doi.org/10.1016/j.inffus.2017.12.002

[22] Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, and Xi Peng. 2020. Partially view-aligned clustering. *Advances in Neural Information Processing Systems* 33 (2020), 2892–2902.

[23] Leonid V Kantorovich. 1942. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, Vol. 37. 199–201.

[24] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. 2014. Partial multi-view clustering. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. 1968–1974.

[25] Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. 2015. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2750–2756.

[26] Yingming Li, Ming Yang, and Zhongfei Zhang. 2018. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering* 31, 10 (2018), 1863–1883.

[27] Zhenglai Li, Chang Tang, Xinwang Liu, Xiao Zheng, Guanghui Yue, Wei Zhang, and En Zhu. 2021. Consensus graph learning for multi-view clustering. *IEEE Transactions on Multimedia* (2021).

[28] Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, and Zhaohua Yang. 2019. Deep adversarial multi-view clustering network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 2952–2958.

[29] Jing Liu, Yu Jiang, Zechao Li, Zhi-Hua Zhou, and Hanqing Lu. 2014. Partially shared latent factor learning with multiview data. *IEEE transactions on neural networks and learning systems* 26, 6 (2014), 1233–1246.

[30] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. 2013. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM international conference on data mining*. SIAM, 252–260.

[31] Xinwang Liu, Yong Dou, Jianping Yin, Lei Wang, and En Zhu. 2016. Multiple kernel k-means clustering with matrix-induced regularization. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 1888–1894.

[32] Yan Luo, Chongyang Zhang, Muming Zhao, Hao Zhou, and Jun Sun. 2020. Where, What, Whether: Multi-modal learning meets pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14065–14073.

[33] Nabil El Malki, Robin Cugny, Olivier Teste, and Franck Ravat. 2020. DECWA: Density-Based Clustering using Wasserstein Distance. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 2005–2008. https://doi.org/10.1145/3340531.3412125

[34] Brian McFee, Gert Lanckriet, and Tony Jebara. 2011. Learning Multi-modal Similarity. *Journal of machine learning research* 12, 2 (2011).

[35] Facundo Mémoli. 2011. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics* 11, 4 (2011), 417–487.

[36] Facundo Mémoli. 2011. A spectral notion of Gromov–Wasserstein distance and related methods. *Applied and Computational Harmonic Analysis* 30, 3 (2011), 363–401.

[37] Facundo Mémoli and Guillermo Sapiro. 2004. Comparing point clouds. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. 32–40.

[38] Andrew Y Ng, Michael I Jordan, and Yair Weiss. 2001. On spectral clustering: analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. 849–856.

[39] Feiping Nie, Guohao Cai, and Xuelong Li. 2017. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 2408–2414.

[40] Feiping Nie, Jing Li, and Xuelong Li. 2017. Self-weighted multiview clustering with multiple graphs. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2564–2570.

[41] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. 2019. COMIC: Multi-view clustering without parameter selection. In *International Conference on Machine Learning*. 5092–5101.

[42] Gabriel Peyré, Marco Cuturi, and Justin Solomon. 2016. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*. PMLR, 2664–2672.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[45] Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. 2021. Linear-time gromov wasserstein distances using low rank couplings and costs. *arXiv preprint arXiv:2106.01128* (2021).

[46] Karl-Theodor Sturm. 2006. On the geometry of metric measure spaces. *Acta mathematica* 196, 1 (2006), 65–131.

[47] Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary. 2019. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*. PMLR, 6275–6284.

[48] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. 2018. Wasserstein Auto-Encoders. In *International Conference on Learning Representations*.

[49] Grigorios Tzortzis and Aristidis Likas. 2012. Kernel-based weighted multi-view clustering. In *2012 IEEE 12th international conference on data mining*. IEEE, 675–684.

[50] Javier Vía, Ignacio Santamaría, and Jesús Pérez. 2007. A learning algorithm for adaptive canonical correlation analysis of several data sets. *Neural Networks* 20, 1 (2007), 139–152.

[51] Cédric Villani. 2008. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.

[52] Huahua Wang and Arindam Banerjee. 2014. Bregman alternating direction method of multipliers. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*. 2816–2824.

[53] Hua Wang, Heng Huang, and Chris Ding. 2011. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 279–284.

[54] Hao Wang, Yan Yang, and Bing Liu. 2019. GMC: Graph-based multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering* 32, 6 (2019), 1116–1129.

[55] Siwei Wang, Xinwang Liu, En Zhu, Chang Tang, Jiyuan Liu, Jingtao Hu, Jingyuan Xia, and Jianping Yin. 2019. Multi-view clustering via late fusion alignment maximization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 3778–3784.

[56] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On Deep Multi-View Representation Learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37* (Lille, France) *(ICML'15)*. JMLR.org, 1083–1092.

[57] Xu Wang, Dezhong Peng, Peng Hu, and Yongsheng Sang. 2019. Adversarial correlated autoencoder for unsupervised multi-view representation learning. *Knowledge-Based Systems* 168 (2019), 109–120.

[58] Jie Wen, Zhihao Wu, Zheng Zhang, Lunke Fei, Bob Zhang, and Yong Xu. 2021. *Structural Deep Incomplete Multi-View Clustering Network*. Association for Computing Machinery, New York, NY, USA, 3538–3542. https://doi.org/10.1145/3459637.3482192

[59] Jie Wen, Yong Xu, and Hong Liu. 2018. Incomplete multiview spectral clustering with adaptive graph learning. *IEEE transactions on cybernetics* 50, 4 (2018), 1418–1429.

[60] Jie Wen, Zheng Zhang, Zhao Zhang, Lei Zhu, Lunke Fei, Bob Zhang, and Yong Xu. 2021. Unified tensor framework for incomplete multi-view clustering and missing-view inferring. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 10273–10281.

[61] Min Xiao and Yuhong Guo. 2012. Multi-view adaboost for multilingual subjectivity analysis. In *Proceedings of COLING 2012*. 2851–2866.

[62] Yuan Xie, Bingqian Lin, Yanyun Qu, Cuihua Li, Wensheng Zhang, Lizhuang Ma, Yonggang Wen, and Dacheng Tao. 2020. Joint deep multi-view learning for image clustering. *IEEE Transactions on Knowledge and Data Engineering* 33, 11 (2020), 3594–3606.

[63] Chang Xu, Dacheng Tao, and Chao Xu. 2015. Multi-view learning with incomplete views. *IEEE Transactions on Image Processing* 24, 12 (2015), 5812–5825.

[64] Hongtengl Xu. 2020. Gromov-Wasserstein factorization models for graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 6478–6485.

[65] Hongteng Xu, Dixin Luo, and Lawrence Carin. 2019. Scalable gromov-wasserstein learning for graph partitioning and matching. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 3052–3062.

[66] Hongteng Xu, Dixin Luo, Ricardo Henao, Svati Shah, and Lawrence Carin. 2020. Learning autoencoders with relational regularization. In *International Conference on Machine Learning*. PMLR, 10576–10586.

[67] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. 2019. Gromov-wasserstein learning for graph matching and node embedding. In *International conference on machine learning*. PMLR, 6932–6941.

[68] TengQi Ye, Tianchun Wang, Kevin McGuinness, Yu Guo, and Cathal Gurrin. 2016. Learning multiple views with orthogonal denoising autoencoders. In *International Conference on Multimedia Modeling*. Springer, 313–324.

[69] Jun Yin and Shiliang Sun. 2022. Incomplete multi-view clustering with cosine similarity. *Pattern Recognition* 123 (2022), 108371.

[70] Hong Yu, Jia Tang, Guoyin Wang, and Xinbo Gao. 2021. A Novel Multi-View Clustering Method for Unknown Mapping Relationships Between Cross-View Samples. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2075–2083.

[71] Changqing Zhang, Ehsan Adeli, Tao Zhou, Xiaobo Chen, and Dinggang Shen. 2018. Multi-layer multi-view classification for alzheimer's disease diagnosis. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. 4406–4413.

[72] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. 2020. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence* (2020).

[73] Changqing Zhang, Yeqing Liu, and Huazhu Fu. 2019. Ae2-nets: Autoencoder in autoencoder networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2577–2585.

[74] Rui Zhang, Hanghang Tong, Yinglong Xia, and Yada Zhu. 2019. Robust Embedded Deep K-Means Clustering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1181–1190. https://doi.org/10.1145/3357384.3357985

[75] Xiaobo Zhang, Yan Yang, Tianrui Li, Yiling Zhang, Hao Wang, and Hamido Fujita. 2021. CMC: a consensus multi-view clustering model for predicting Alzheimer's disease progression. *Computer Methods and Programs in Biomedicine* 199 (2021), 105895.

[76] Xianchao Zhang, Linlin Zong, Xinyue Liu, and Hong Yu. 2015. Constrained NMF-based multi-view clustering on unmapped data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 3174–3180.

[77] Sihang Zhou, Xinwang Liu, Miaomiao Li, En Zhu, Li Liu, Changwang Zhang, and Jianping Yin. 2019. Multiple kernel clustering with neighbor-kernel subspace segmentation. *IEEE transactions on neural networks and learning systems* 31, 4 (2019), 1351–1362.

[78] Pengfei Zhu, Binyuan Hui, Changqing Zhang, Dawei Du, Longyin Wen, and Qinghua Hu. 2019. Multi-view deep subspace clustering networks. *arXiv preprint arXiv:1908.01978* (2019).

[79] Wenjie Zhu, Bo Peng, and Chunchun Chen. 2021. *Self-Supervised Embedding for Subspace Clustering*. Association for Computing Machinery, New York, NY, USA, 3687–3691. https://doi.org/10.1145/3459637.3482178

[80] Xiaofeng Zhu, Shichao Zhang, Wei He, Rongyao Hu, Cong Lei, and Pengfei Zhu. 2018. One-step multi-view spectral clustering. *IEEE Transactions on Knowledge and Data Engineering* 31, 10 (2018), 2022–2034.

[81] Linlin Zong, Faqiang Miao, Xianchao Zhang, and Bo Xu 0008. 2020. Multimodal Clustering via Deep Commonness and Uniqueness Mining. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 2357–2360. https://doi.org/10.1145/3340531.3412103