Deep Incomplete Multi-view Clustering with Cross-view Partial Sample and Prototype Alignment

Jiaqi Jin[1], Siwei Wang[1], Zhibin Dong[1], Xinwang Liu[1,*], En Zhu[1,*]

[1]School of Computer, National University of Defense Technology, Changsha, China

{jinjiaqi, wangsiwei13, dzb20, xinwangliu, enzhu}@nudt.edu.cn

Abstract

The success of existing multi-view clustering relies on the assumption of sample integrity across multiple views. However, in real-world scenarios, samples of multi-view are partially available due to data corruption or sensor failure, which leads to incomplete multi-view clustering study (IMVC). Although several attempts have been proposed to address IMVC, they suffer from the following drawbacks: i) Existing methods mainly adopt cross-view contrastive learning forcing the representations of each sample across views to be exactly the same, which might ignore view discrepancy and flexibility in representations; ii) Due to the absence of non-observed samples across multiple views, the obtained prototypes of clusters might be unaligned and biased, leading to incorrect fusion. To address the above issues, we propose a Cross-view Partial Sample and Prototype Alignment Network (CPSPAN) for Deep Incomplete Multi-view Clustering. Firstly, unlike existing contrastive-based methods, we adopt pair-observed data alignment as 'proxy supervised signals' to guide instance-to-instance correspondence construction among views. Then, regarding of the shifted prototypes in IMVC, we further propose a prototype alignment module to achieve incomplete distribution calibration across views. Extensive experimental results showcase the effectiveness of our proposed modules, attaining noteworthy performance improvements when compared to existing IMVC competitors on benchmark datasets.

1. Introduction

In modern society, data collected for real-world applications usually stems from different domains, sensors or feature extractors, which gives rise to multi-view learning in literature [2, 44]. For instance, an autonomous car may have diverse sensors, and a movie is typically made up of images and audio. As an important paradigm of multi-view

learning, multi-view clustering (MVC) [10,20,21,24,40,47] divides data by exploiting the consistent and complementary information across multiple views. The success of existing multi-view clustering methods heavily relies on the fully-available data assumption. However, in practical applications, some views of instances are only partially available due to unstable sensors and damaged storage media. When some views are missing [9], the natural alignment property of same instances across multiple views is destroyed, which may result in insufficient mining of complementary and consistent information. To handle the incompleteness issue, many incomplete multi-view clustering algorithms (IMVC) [15,31,38] with satisfactory performance have been proposed. Typical strategies are mainly based on matrix decomposition, incomplete multiple kernel learning and graph-based methods. Learning more discriminative consensus representations with incomplete view information is crucial to achieve better incomplete multi-view clustering performance. However, conventional IMVC methods are based on raw features and therefore, the performance heavily relies on the feature quality.

As deep neural networks [5,8,16,17,25,43] have demonstrated superior performance in learning high-level representations, deep learning has become prevalent in various fields of computer vision and pattern classification. To this end, researchers have explored combining deep neural networks [26, 35] and conventional IMVC methods to improve clustering performance, and the resulting clustering method is called Deep Incomplete Multi-View Clustering (DIMVC) [11, 30, 37, 39, 42]. Most existing DIMVC methods adopt the principle of contrastive learning, treating different views of the same sample as positive pairs and their representations should be consistent. Such algorithms ignore the cross-view alignment correlation of samples and force instances of different views with unified representation, which may destroy the flexibility and variety of representations. We argue that the essence of IMVC task lies in discovering structural correspondence between different views, rather than rigidly and simply enforcing uniform representations across each view. In fact, IMVC can be re-

---
*Corresponding author

garded as a special case of 'partially-aligned' multi-view setting, where the pair-observed data provides supervised instance-alignment signals.

Moreover, as shown in the Fig. 1, the distribution learned from the incomplete multi-view data can be biased due to inadequate multi-view data. Specifically, during the clustering task, flexible representations may cause prototypes of each cluster to shift and become biased, which we refer to as the Prototype-Shifted Problem (termed PSP). Such a problem has been demonstrated in the Anchor-Unaligned Problem [32] for complete multi-view data, and undoubtedly has more essential impact on incomplete multi-view data. At the same time, contrastive-based DIMVC methods neglect this issue and do not explore relationships among different instances within the same view, which may further aggravate PSP. Therefore, it is necessary to match the relationship between the prototypes among views and perform clustering task accordingly.

To address the aforementioned issues, we propose a novel approach termed **C**ross-view **P**artial **S**ample and **P**rototype **A**lignment **N**etwork (CPSPAN) for Deep Incomplete Multi-view Clustering to perform cross-view partial sample alignment and solve the prototype-shifted problem. The framework of CPSPAN is illustrated in Fig. 3. In detail, different from the contrastive learning mode, the cross-view instance alignment module establishes the view-to-view correspondence of samples through the pair-observed data in Fig. 2 between each pair-wise views, so as to mine the structural information between views. Afterwards, to address the prototype-shifted problem in incomplete scenario, the prototype alignment module takes one view's prototype set as anchors, and solves the permutation matrix between the two sets of prototypes, thereby establishing prototype-to-prototype correspondence based on optimal transport theory. Since prototypes are obtained based on samples, this module not only calibrates correspondences between cross-view shifted prototypes but also encodes the relationships between within-view samples. Ultimately, since our model is imputation-free upfront, in order to align the embeddings between views before finally performing feature fusion and clustering, we build cross-view structural relationship transfer for missing item imputations.

We summarize the major contributions of our work as follows,

- We propose a novel deep network to handle IMVC task, termed as CPSPAN. Differ from existing multi-view contrastive learning manner, we considers the IMVC from a novel insight with **partially-aligned** setting. To this end, CPSPAN optimal maximizes matching alignment between paired-observed data and construct cross-view intersection.
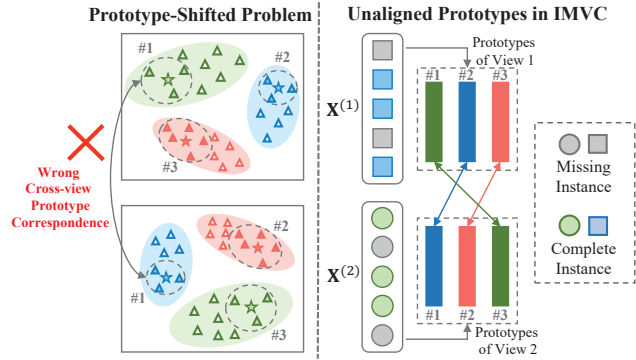

Figure 1. An example illustration of shifted prototype across multi-view caused by incomplete setting. With different missing status, the prototypes learned by incomplete multi-view data may be shifted and leads to wrong correspondences.

- In order to solve the Prototype-Shifted Problem caused by incomplete information, CPSPAN proposes to further align the prototype sets between different views, so as to mine consistent cross-view structural information.

- Extensive experiments have clearly demonstrated the effectiveness of the proposed cross-view partial sample and prototype alignment modules and the superiority over both conventional and deep SOTA methods.

2. Related Work

This section briefly describes the latest research progress of Incomplete Multi-View Clustering (IMVC) and Deep Incomplete Multi-View Clustering (DIMVC).

2.1. Incomplete Multi-view Clustering

The incomplete multi-view clustering problem has attracted extensive attention of researchers in recent years. Existing methods for incomplete multi-view clustering mostly fall into three categories: (1) Matrix factorization based IMVC [6, 9]. Non-negative matrix factorization is dedicated to mining consensus representations of incomplete multi-view data. DAIMC [6] is based on semi-NMF to obtain consensus information among multiple views. Partial multi-view clustering (PVC) [9] exploits NMF to mine latent information of incomplete data. (2) Kernel learning based IMVC [14, 15]. Kernel-based methods aim to recover the kernel matrix for incomplete multi-view data. ONKC [14] enhances representability of the optimal kernel and strengthens negotiation between kernel learning and clustering. (3) Graph learning based IMVC [48, 49]. IMG [49] extends PVC by exploring rich multi-view global structural information. Nevertheless, traditional IMVC methods have limitation in exploring the clustering-friendly repre-
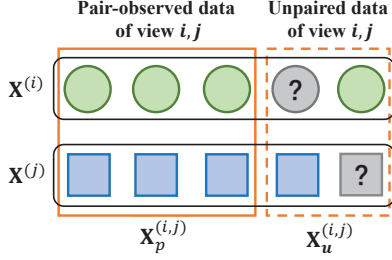
Figure 2. Definition of pair-observed data and unpaired data between view i and j. Instances that are complete on both view i and j are referred as pair-observed data.

sentation of high-dimensional complex data, and they heavily rely on the quality of raw features.

2.2. Deep Incomplete Multi-view Clustering

With the development of deep learning, researchers have realized the important role of representation learning in clustering task. The high-level feature mining capability of deep neural networks has attracted the attention of researchers, resulting in deep models being widely used in incomplete multi-view clustering tasks. The existing DIMVC methods can be divided into the following three categories: (1) Autoencoders-based methods. By extracting feature of data, the deep autoencoder can learn the consistent representation to pad missing data and obtain impressive clustering performance. Typical methods are [50], [36], [30] and [12]. (2) GANs-based methods. This kind of approaches explore the mutual representation between multiple views by generating adversarial networks and directly producing missing data [1, 3, 27–29]. (3) GCNs-based methods. GCNs-based methods explore common representations of multiple views through structural information among them [33, 34]. (4) Contrastive-based methods. The representative work is [11]. This method completes the missing data and explores the congruous representation through the comparison between multiple views. However, contrastive-based methods destroy the flexibility and diversity of representations and lose complementary information in incomplete multi-view clustering.

3. Proposed Method

In this section, we first elaborate on the motivations of our work. Then we comprehensively introduce the proposed CPSPAN which consists of three jointly learning modules, namely, incomplete multi-view representation learning, cross-view partial sample alignment and shifted prototype alignment module.

Notations. Formally, a multi-view incomplete dataset $\{\mathbf{X}^{(v)} \in \mathbb{R}^{N \times D_v}\}_{v=1}^V$ consists of N instances of V views.

We denote the samples with complete data as a set $\{\tilde{\mathbf{X}}^{(v)} \in \mathbb{R}^{N_v \times D_v}\}_{v=1}^V$, where N_v denotes the number of complete instance in v-th view. As shown in Fig. 2, we utilize $\mathbf{X}_p^{(i,j)}$ and $\mathbf{X}_u^{(i,j)}$ to denote the pair-observed and unpaired instances respectively. K is the number of clusters.

3.1. Representation Learning Module and Prototype Set Generation

Incomplete Multi-view Representation Learning. In multi-view clustering tasks, deep autoencoders have been widely used to extract high-level representations of raw features [4, 12, 22, 41, 45]. Therefore, we equip each view with an auto-encoder to learn the clustering-friendly features of the respective view. To avoid inaccurate imputation or padding negatively impacting representation learning, we only train with the complete instance in each view. In order to avoid instance misalignment between views caused by missing data, we use resampling method to complement each view by sampling observed instances on each view. For the v-th view, an encoder-decoder pair, denoted as E_v and D_v, is utilized to obtain the high-level embedding $\mathbf{H}^{(v)}$. Specifically, the feature extraction process of E_v is expressed as $E_v(\tilde{\mathbf{X}}^{(v)}; \theta^v) : \tilde{\mathbf{X}}^{(v)} \in \mathbb{R}^{N_v \times D_v} \rightarrow \mathbf{H}^{(v)} \in \mathbb{R}^{N_v \times d}$. The process of feature reconstruction of $\mathbf{H}^{(v)}$ by the decoder D_v is formulated as $D_v(\mathbf{H}^{(v)}; \phi^v) : \mathbf{H}^{(v)} \in \mathbb{R}^{N_v \times d} \rightarrow \hat{\mathbf{X}}^{(v)} \in \mathbb{R}^{N_v \times D_v}$, where d is the dimension of the embedding on all views, θ^v and ϕ^v are the learnable parameters of autoencoder for the v-th view. Autoencoders are utilized to project the features of all views into the same dimensionality, which enables measuring correspondence between prototypes of different views in the same dimension. The reconstruction loss for all views is expressed as

$$
\begin{aligned}
\mathcal{L}_{rec} &= \sum_{v=1}^V \left\| \tilde{\mathbf{X}}^{(v)} - D_v(\mathbf{H}^v) \right\|_F^2 \\
&= \sum_{v=1}^V \sum_{i=1}^{N_v} \left\| \tilde{\mathbf{x}}_i^{(v)} - D_v(E_v(\tilde{\mathbf{x}}_i^{(v)})) \right\|_2^2,
\end{aligned}
\tag{1}
$$

our model is first pre-trained in the above way, so far it is still independent on each view.

Prototype Set Generation. In the d-dimensional feature space of each view, the prototype set of each view $\{\mathbf{C}^{(v)}\}_{v=1}^V$ can be obtained by the following objective:

$$
\min_{\mathbf{C}^{(v)}} \sum_{i=1}^{N_v} \sum_{j=1}^K \left\| \mathbf{h}_i^{(v)} - \mathbf{c}_j^{(v)} \right\|_2^2
\tag{2}
$$

where $\mathbf{C}^{(v)} \in \mathbb{R}^{K \times d}$. The subsequent cross-view partial sample alignment and shifted prototype alignment are performed between the features generated by the E_v and D_v of each view and the prototype set $\mathbf{C}^v$ generated by Eq. (2).
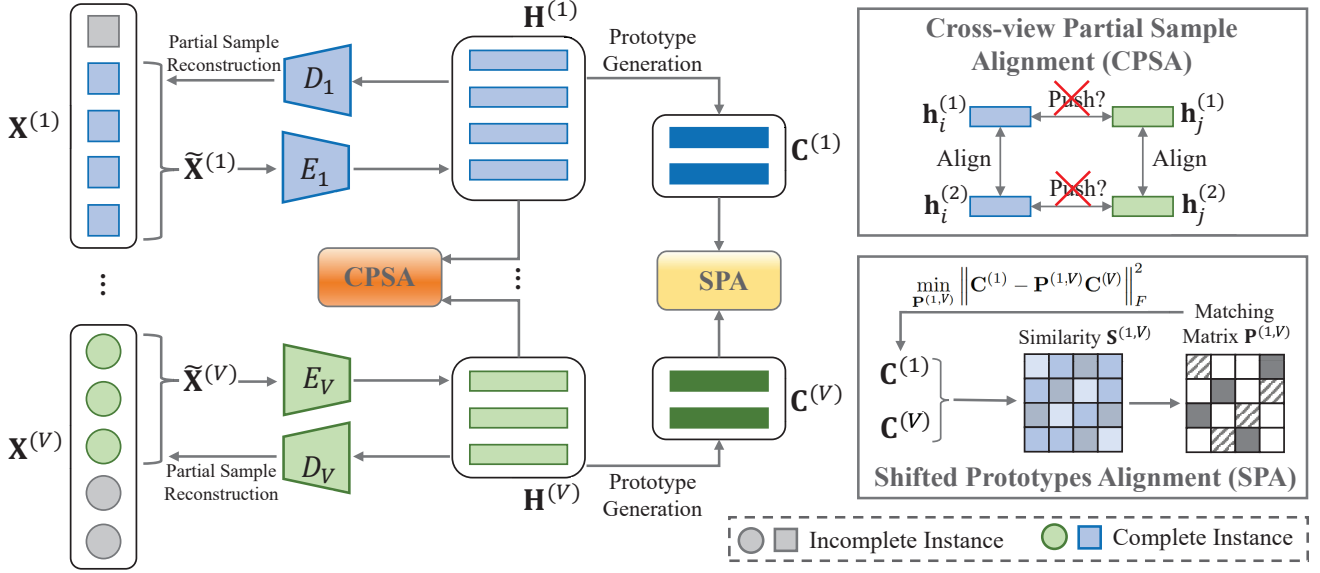
Figure 3. The framework of our proposed CPSPAN. CPSPAN consists of three modules: incomplete multi-view representation learning module, cross-view partial sample alignment module (CPSA) and shifted prototype alignment (SPA) module. Specifically, CPSA performs view-to-view alignment on learned representations of pair-observed instances. SPA explores the optimal matching correspondence between prototypes so as to align them. Finally, we use the structure embedding imputation strategy to fill in the missing embedding. Ultimately, we concatenate the complete and filled embedding and then perform kmeans to get the final result.

3.2. Cross-view Partial Sample Alignment

In the context of missing data, the multi-view clustering method based on contrastive learning leverages V representations of the same sample in different views as positive samples and views of other samples as negative samples. By doing so, the method forces the representations of different views corresponding to the same sample to converge, ultimately enabling the learning of consistent information across views. The commonly used loss function associated with this method is

$$\ell_i^{(v)} = -\sum_{\substack{r=1, \\ r \neq v}}^{V} \log \frac{\exp\left(h_i^{(v)} h_i^{(r)} / \tau\right)}{\sum_{\substack{t=1, \\ t \neq v}}^{V} \sum_{\substack{j=1, \\ j \neq i}}^{N} \exp\left(h_i^{(v)} h_j^{(t)} / \tau\right)}, \quad (3)$$

$$\mathcal{L}_{cl} = \frac{1}{N} \sum_{v=1}^{V} \sum_{i=1}^{N} \ell_i^{(v)}, \quad (4)$$

where $h_i^{(r)}$ is the positive instance of the i-th sample in other view. τ is the temperature parameter to control the softness. After the optimization of the loss function in Eq. (4), the representation across views are forced to be exactly the same.

Different from previous methods based on the principle of contrastive learning, our cross-view instance alignment module can learn more flexible representations. The representations learned by our module are more flexible for two reasons. One is that the module uses the cosine similarity to measure the difference between paired data between two views in the feature space, rather than using the vector inner product or the Euclidean distance of the vector in the feature space, which is different from the clustering that pays more attention to mining the structural information of data. On the other hand, our method only restricts the representation of samples in different views to achieve the highest matching degree, and does not enforce the matching degree with other samples to be 0. This alignment approach avoids stretching samples that originally belong to the same cluster and is better suited to mining structural information.

Note that $\mathbf{X}_p^{(i,j)}$ represents the pair-observed data in view i and j. The number of instances in $\mathbf{X}_p^{(i,j)}$ is $N_{(i,j)}$, and the corresponding paired embedding of $\mathbf{X}_p^{(i,j)}$ in the feature space is $\mathbf{H}_p^{(i,j)}$. We define the difference between the p-th instance in view i and the q-th sample in view j in $\mathbf{H}_p^{(i,j)}$ as

$$\mathbf{S}_{p,q}^{(i,j)} = \frac{(\mathbf{h}_p^{(i,j)})^T \mathbf{h}_q^{(i,j)}}{\|\mathbf{h}_p^{(i,j)}\| \|\mathbf{h}_q^{(i,j)}\|}, \quad (5)$$

where $\mathbf{S}^{(i,j)} \in \mathbb{R}^{N_{(i,j)} \times N_{(i,j)}}$ is the similarity matrix between paired embeddings in view i and j.

As mentioned above, we only restrict all elements on the diagonal of $S^{(i,j)}$ to be 1, and do not impose constraints on other elements of $S^{(i,j)}$. We impose this constraint on paired embedding of each pair-wise views, so the loss func-

tion of our cross-view instance alignment module can be defined as

$$\mathcal{L}_{ia} = \sum_{\substack{i=1,\cdots,V \\ j=i+1}} \left\| diag(\mathbf{S}^{(i,j)}) - \mathbf{1}_{N_{(i,j)}} \right\|_2^2, \qquad (6)$$

3.3. Shifted Prototype Alignment

By matching the prototype-to-prototype correspondence between each pairwise views, this module calibrates the relationship between samples in a view and the relationship between prototypes on different views, thereby solving the prototype-shifted problem and further improving the clustering performance.

Formally, the crucial step of this module is to obtain the permutation matrix $\mathbf{P} \in \mathbb{R}^{K \times K}$ between prototypes of different views. To achieve prototype alignment, we optimize P as an integer linear programming problem with the objective of achieving an optimal match between two prototype sets. Formally,

$$\begin{aligned}
\arg\min_{\mathbf{P}} \quad & Tr(\mathbf{DP}^T), \\
\text{s.t.} \quad & P_{ij} \geq 0, \forall (i,j), \\
& \mathbf{P1} = \mathbf{1}, \\
& \mathbf{P}^T\mathbf{1} = \mathbf{1},
\end{aligned} \qquad (7)$$

where $\mathbf{P}$ is a square binary matrix that has exactly one entry of 1 in each row and column, and 0 elsewhere. We denote the permutation matrix between views i and j by $\mathbf{P}^{(i,j)}$. Taking the prototypes $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$ on view 1 and view 2 as an example, if $\mathbf{C}^{(2)}$ is aligned to $\mathbf{C}^{(1)}$, after finding the best matching flow $\mathbf{P}^{(1,2)}$ between the two, use the permutation matrix $\mathbf{P}^{(1,2)}$ to left-multiply $\mathbf{C}^{(2)}$ to align to $\mathbf{C}^{(1)}$. The optimal formula for solving $\mathbf{P}^{(1,2)}$ is

$$\min_{\mathbf{P}^{(1,2)}} \left\| \mathbf{C}^{(1)} - \mathbf{P}^{(1,2)}\mathbf{C}^{(2)} \right\|_F^2. \qquad (8)$$

In order to make the optimization of the permutation matrix suitable for the derivation mechanism in deep neural networks, we follow the method in PVC [7] and use the differentiable substitution of the Hungarian algorithm to establish the correspondence of prototype sets. Formally,

$$\Omega_1 = ReLU(\mathbf{P}^{(1,2)}), \qquad (9)$$

$$\Omega_2 = \mathbf{P}^{(1,2)} - \frac{1}{n}(\mathbf{P}^{(1,2)}\mathbf{1} - \mathbf{1})\mathbf{1}^T, \qquad (10)$$

$$\Omega_3 = \mathbf{P}^{(1,2)} - \frac{1}{n}\mathbf{1}(\mathbf{1}^T\mathbf{P}^{(1,2)} - \mathbf{1}^T), \qquad (11)$$

where Ω_1, Ω_2 and Ω_3 project the permutation matrix $\mathbf{P}^{(1,2)}$ to the three constraint sets.
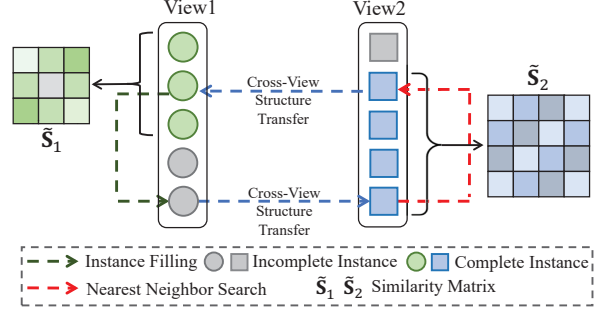

Figure 4. Visualization of Structure Embedding Imputation Strategy. We build cross-view structure transfer to impute instances with their close neighbours in latent space.

Following the above optimization process, we can obtain the permutation matrix and optimally match the prototype sets between each pairwise views, and the corresponding loss function is as follows

$$\mathcal{L}_{pa} = \sum_{\substack{i=1,\cdots,V \\ j=i+1}} \left\| \mathbf{C}^{(i)} - \mathbf{P}^{(i,j)}\mathbf{C}^{(j)} \right\|_F^2. \qquad (12)$$

3.4. Structure Embedding Imputation Strategy

Since our model does not perform imputation during training, in order to align features between views, we propose a structure embedding imputation strategy to pad the missing embedding. Fig. 4 is an example of our imputation strategy. Specifically, first we calculate the similarity matrix based on all complete embeddings in the two views respectively. For a missing feature in view 1, we find the nearest neighbor to its embedding in view 2, and then directly fill the missing feature with that neighbor's embedding in view 1.

3.5. The Objective Function

With the above definitions, we propose the following objective function:

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha\mathcal{L}_{ia} + \beta\mathcal{L}_{pa}, \qquad (13)$$

where $\mathcal{L}_{rec}$, $\mathcal{L}_{ia}$ and $\mathcal{L}_{pa}$ are within-view reconstruction loss for observable instances, cross-view partial sample alignment loss and shifted prototype alignment loss, respectively. In the experiments, we simply set both balance coefficients to 0.001.

4. Experiments

4.1. Experimental Setting

Datasets and Evaluation Metrics. We conduct experiments on five benchmark multi-view datasets, namely

Caltech101-7, HandWritten, ALOI-100, YouTubeFace10 and EMNIST. Details of these data are shown in Table 1 below. In this experiment, we set the missing rate of the each dataset to [0.1, 0.3, 0.5, 0.7]. To evaluate the validity of the experiments, we used three widely used evaluation criteria including accuracy (ACC), normalized mutual information (NMI), and F-mea.

Table 1. Incomplete multi-view datasets in experiments.

Dataset	Samples	Clusters	Views	Dimensionality
Caltech101-7	1400	7	5	1984/512/928/254/40
HandWritten	2000	10	6	216/76/64/6/240/47
ALOI-100	10800	100	4	77/13/64/125
YouTubeFace10	38654	10	4	944/576/512/640
EMNIST	280000	10	4	944/576/512/640

Baseline Methods. In experiments, our proposed algorithm is compared with seven state-of-the-art Incomplete multi-view clustering methods. **Best Single-view Spectral Clustering (BSV)** [18] fills in missing data by mean on single view data. **PIC** [13] strives to find a consensus feature matrix to complement incomplete multi-view data. **AWP** [19] complements the missing data between multiple views through the structural information between multiple views. **CPM-Nets** [46] transforms the multiview representation learning task into a degenerate process to achieve consistency and completion between multiple views. **COMPLETER** [12] learns consistent representations through contrastive learning and maximizes the mutual information between multiple views. **DCP** [11] learns consistent representations and complements information by maximizing mutual information and minimizing cross-entropy. **DSIMVC** [23] dynamically complements the missing views from the learned semantic neighbors, solving the multi-view missing problem.

Implementation. Tne proposed CPSPAN is trained with the PyTorch platform. In the experiments, we set the pre-training batchsize to 256, the pre-training epoch to 200, and alignment epoch is 50. For all datasets, the autoencoders for all views are implemented by MLPs with the same structure, and the feature dimension in the embedding space is set to 10. The activation function is ReLU. We adopt Adam to optimize the deep model, the learning rate in the pre-training stage is set to 0.0005, and the learning rate in the alignment stage is set to 0.0001.

4.2. Comparison with State-of-The-Arts

In Table 2 of the paper, our proposed algorithm is compared with seven state-of-the-art methods on three metrics. From the above chart, we have the following observations:

- Our proposed model significantly outperforms other incomplete multi-view clustering algorithms on all datasets in terms of all metrics. Under the 0.1 missing rate setting, for all the datasets, our proposed
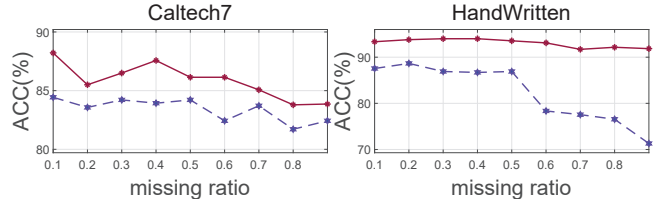

Figure 5. Flexibility verification of the cross-view instance alignment module. The red lines are the flexible features learned through our proposed CPSPAN, and the blue lines are the features learned after changing the loss function to contrastive loss.

method are all significantly higher than the second best approaches, and outperform the second best by 5.0%, 2.9%, 0.6%, 3.3% and 1.5%, respectively. This can demonstrate the superiority of our method on all datasets. Moreover, CPSPAN demonstrates greater stability as the missing rate varies. For instance, when the missing rate ranges from 0.1 to 0.7 on HandWritten, the performance only exhibits a modest 1.7% decrease. Similarly, on the EMNIST, where the variation in the method's performance is greatest, the drop is only 6.4%.

- It can be seen that the CPSPAN can still achieve competitive effect with a high missing rate in the dataset. When the miss rate is 0.7, the proposed method is 7.9%, 5.9%, 1.3%, 7.3% and 15.4% higher than the method with the second highest ACC in the five datasets, respectively. It can be seen that our proposed method can still achieve impressive results in the case of a relatively high missing rate, which demonstrates the stability of CPSPAN.

In conclusion, the above experiments have verified the effectiveness and stability of our proposed CPSPAN and it is pivotal to conduct pair-ovserved sample alignment and prototype alignment for IMVC task.

4.3. Flexibility Validation of Cross-view Partial Sample Alignment

As mentioned in Section 1, our cross-view partial sample alignment strategy is more flexible than the commonly used contrastive-based methods, and can learn more clustering-friendly representation and mine better structural information of data. To verify this, we directly replace our pair-observed sample alignment loss with contrastive loss in Eq. (4). We validate our idea on both Caltech101-7 and HandWritten. From the results in Fig. 5, it can be seen that our method outperforms contrastive-based approach at any missing rate, which indicates that our cross-view partial sample alignment module is more conducive to clustering representation.

Table 2. The clustering performance comparisons on five benchmark datasets with varying missing rates. The best results are highlighted in bold, while the second best results are underlined. 'O/M' denotes the out-of-memory failure due to the size of the dataset.

	Missing rates	0.1			0.3			0.5			0.7		
	Metrics	ACC	NMI	F-mea	ACC	NMI	F-mea	ACC	NMI	F-mea	ACC	NMI	F-mea
Caltech101-7	BSV	0.228	0.036	0.246	0.284	0.250	0.297	0.265	0.210	0.281	0.245	0.166	0.268
	PIC	0.656	0.592	0.649	0.653	0.619	0.654	0.646	0.605	0.641	0.644	0.607	0.647
	AWP	0.779	0.735	0.724	0.724	0.677	0.677	0.661	0.667	0.636	0.772	0.691	0.695
	CPM-Nets	0.832	0.734	0.811	0.754	0.652	0.721	0.696	0.575	0.693	0.611	0.499	0.605
	COMPLETER	0.528	0.573	0.499	0.500	0.533	0.485	0.534	0.557	0.535	0.596	0.587	0.585
	DCP	0.578	0.584	0.552	0.454	0.518	0.424	0.487	0.548	0.443	0.172	0.059	0.092
	DSIMVC	0.649	0.565	0.642	0.629	0.537	0.622	0.554	0.471	0.549	0.419	0.343	0.391
	CPSPAN	**0.882**	**0.793**	**0.878**	**0.865**	**0.768**	**0.861**	**0.861**	**0.775**	**0.856**	**0.851**	**0.760**	**0.843**
HandWritten	BSV	0.584	0.614	0.522	0.532	0.553	0.446	0.482	0.503	0.368	0.437	0.454	0.292
	PIC	0.779	0.797	0.720	0.781	0.792	0.749	0.800	0.833	0.783	0.802	0.822	0.782
	AWP	0.708	0.816	0.719	0.867	0.880	0.839	0.860	0.874	0.834	0.858	**0.858**	0.823
	CPM-Nets	0.905	0.827	0.905	0.840	0.754	0.851	0.753	0.627	0.770	0.663	0.578	0.705
	COMPLETER	0.847	0.852	0.833	0.610	0.679	0.615	0.481	0.614	0.485	0.730	0.730	0.727
	DCP	0.682	0.793	0.644	0.769	0.788	0.765	0.670	0.738	0.652	0.708	0.742	0.686
	DSIMVC	0.797	0.749	0.794	0.789	0.731	0.785	0.743	0.687	0.738	0.656	0.624	0.643
	CPSPAN	**0.934**	**0.880**	**0.933**	**0.940**	**0.888**	**0.940**	**0.936**	**0.879**	**0.935**	**0.917**	0.847	**0.917**
ALOI-100	BSV	0.412	0.741	0.268	0.369	0.741	0.237	0.322	0.740	0.236	0.290	0.742	0.238
	PIC	0.698	0.608	0.436	0.633	0.345	0.242	0.541	0.465	0.365	0.351	0.355	0.235
	AWP	0.706	0.785	0.579	0.679	0.785	0.568	0.674	0.757	0.538	0.673	0.757	0.538
	CPM-Nets	0.242	0.529	0.196	0.153	0.430	0.139	0.139	0.346	0.121	0.074	0.259	0.059
	COMPLETER	0.169	0.421	0.143	0.174	0.419	0.164	0.144	0.384	0.135	0.143	0.403	0.173
	DCP	0.272	0.521	0.255	0.230	0.477	0.220	0.227	0.488	0.221	0.213	0.451	0.215
	DSIMVC	0.325	0.621	0.209	0.306	0.586	0.250	0.299	0.572	0.199	0.281	0.552	0.230
	CPSPAN	**0.712**	**0.864**	**0.738**	**0.688**	**0.856**	**0.654**	**0.709**	**0.864**	**0.679**	**0.686**	**0.859**	**0.662**
YouTubeFace10	BSV	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M
	PIC	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M
	AWP	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M
	CPM-Nets	0.772	0.809	0.753	0.740	0.768	0.715	0.698	0.702	0.673	0.657	0.621	0.633
	COMPLETER	0.620	0.642	0.605	0.653	0.677	0.628	0.551	0.563	0.534	0.484	0.501	0.470
	DCP	0.746	0.782	0.725	0.659	0.644	0.617	0.604	0.577	0.586	0.628	0.604	0.572
	DSIMVC	0.730	0.762	0.714	0.716	0.738	0.689	0.683	0.691	0.657	0.664	0.695	0.642
	CPSPAN	**0.805**	**0.839**	**0.795**	**0.794**	**0.802**	**0.734**	**0.750**	**0.767**	**0.729**	**0.741**	**0.765**	**0.691**
EMNIST	BSV	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M
	PIC	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M
	AWP	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M	O/M
	CPM-Nets	0.843	0.812	0.837	0.776	0.749	0.755	0.720	0.688	0.692	0.676	0.621	0.649
	COMPLETER	0.856	0.827	0.835	0.762	0.704	0.711	0.650	0.597	0.618	0.533	0.485	0.509
	DCP	0.733	0.651	0.720	0.675	0.589	0.642	0.692	0.556	0.630	0.584	0.443	0.528
	DSIMVC	0.710	0.672	0.694	0.639	0.572	0.597	0.585	0.524	0.537	0.530	0.472	0.508
	CPSPAN	**0.871**	**0.834**	**0.872**	**0.865**	**0.828**	**0.867**	**0.857**	**0.823**	**0.844**	**0.830**	**0.785**	**0.791**

4.4. Stability Analysis of Imputation Strategy

We analyze the stability of the structure embedding imputation strategy by verifying it with the i-th nearest neighbor method, with i ranging from 1 to 100. The results indicate that although the range of nearest neighbors varies widely, the performance only drops by less than 5% on both datasets. This confirms the stability of our method. More detailed results are provided in the supplementary.

4.5. Parameter Sensitivity Analysis

In order to analyze the effect of the two balance parameters α and β in the loss function on the efficiency of the algorithm, we perform parameter sensitivity analysis on Caltech101-7 and HandWritten. We set the values of both parameters to [1e-3, 1e-2, 1e-1, 1e1, 1e2, 1e3]. From the results in the Fig. 7, we can find that β has a greater impact on Caltech101-7, but both parameters have very little
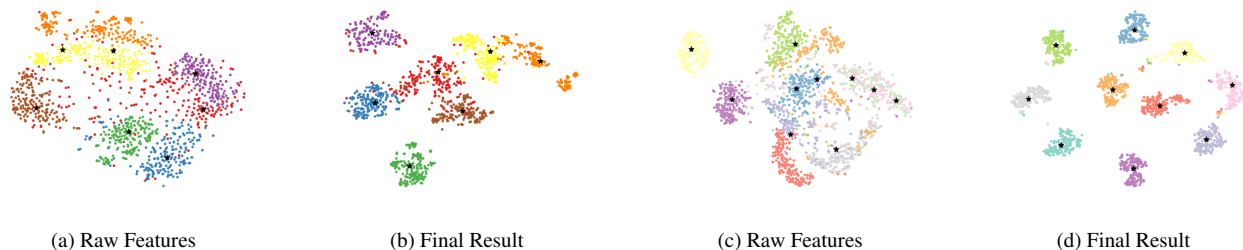
| (a) Raw Features | (b) Final Result | (c) Raw Features | (d) Final Result |

Figure 6. Visualization of Caltech101-7 and HandWritten. '*' indicates the prototype of each cluster.

Table 3. Ablation study on Caltech101-7. We select ACC as the evaluation metric. In the following table, ✓ denotes CPSPAN with the component.

Components			Missing Ratios			
$\mathcal{L}_{rec}$	$\mathcal{L}_{ia}$	$\mathcal{L}_{pa}$	0.1	0.3	0.5	0.7
✓			0.7050	0.6636	0.6286	0.5243
✓	✓		0.8429	0.8243	0.8057	0.8164
✓		✓	0.7757	0.7486	0.7071	0.7093
✓	✓	✓	**0.8821**	**0.8650**	**0.8614**	**0.8507**

Table 4. Ablation study on HandWritten. We select ACC as the evaluation metric. In the following table, ✓ denotes CPSPAN with the component.

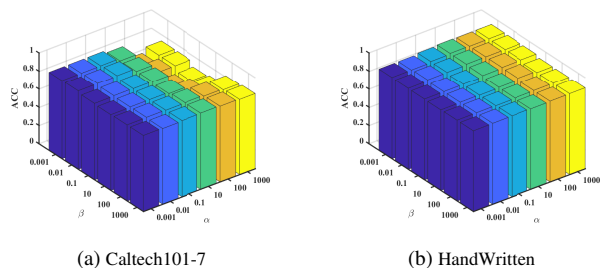| Components			Missing Ratios			
$\mathcal{L}_{rec}$	$\mathcal{L}_{ia}$	$\mathcal{L}_{pa}$	0.1	0.3	0.5	0.7
✓			0.8375	0.8115	0.8035	0.7780
✓	✓		0.9065	0.9035	0.8870	0.8600
✓		✓	0.8885	0.8600	0.8135	0.8075
✓	✓	✓	**0.9335**	**0.9400**	**0.9355**	**0.9170**

| (a) Caltech101-7 | (b) HandWritten |

Figure 7. Sensitivity analysis of α and β for our method over HandWritten and Caltech101-7 (The missing ratio is 0.5).

impact on HandWritten.

4.6. Ablation Study on Modules

The proposed CPSPAN contains three modules. To further verify the importance of each module, we conducted ablation experiments on Caltech101-7 and HandWritten datasets. The results are presented in Tables 3 and 4, with missing rate of [0.1, 0.3, 0.5, 0.7]. Our findings indicate that both cross-view instance alignment and shifted prototype alignment modules contribute significantly to the improved performance of both datasets.

4.7. Visualization

We visualize the clustering effect of CPSPAN on Caltech101-7 and HandWritten under the setting of 0.5 missing rate. Raw features and the final embedding are shown in Fig. 6 respectively. From the figure, we can clearly observe that after the training of our model, the instances from the same cluster become more compact, and the instances of different clusters are separated far away. In addition, the visualization results of the prototypes of each cluster further verify that after the prototype alignment module, the shifted prototype can be re-estimated and accurately calibrated.

5. Conclusion

In this paper, we propose a novel Cross-view Partial Sample and Prototype Alignment framework for incomplete multi-view data termed CPSPAN. Different from other incomplete multi-view methods, CPSPAN uses pair-observed data alignment to guide the correspondence reconstruction between samples. Aiming at the prototype-shifted problem in incomplete multi-view clustering, we also propose a shifted prototype alignment module to calibrate the prototype sets across views. Experiments demonstrate the superiority of our CPSPAN compared with conventional and deep IMVC SOTA methods.

Acknowledgments

References

[1] Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, 1(1):100004, 2021. 3

[2] Lele Fu, Pengfei Lin, Athanasios V Vasilakos, and Shiping Wang. An overview of recent multi-view clustering. *Neurocomputing*, 402:148–161, 2020. 1

[3] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 3

[4] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Ijcai*, pages 1753–1759, 2017. 3

[5] Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 1133–1141. IEEE, 2017. 1

[6] Menglei Hu and Songcan Chen. Doubly aligned incomplete multi-view clustering. *arXiv preprint arXiv:1903.02785*, 2019. 2

[7] Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, and Xi Peng. Partially view-aligned clustering. *Advances in Neural Information Processing Systems*, 33:2892–2902, 2020. 5

[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[9] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014. 1, 2

[10] Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015. 1

[11] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 3, 6

[12] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11174–11183, 2021. 3, 6

[13] Xinwang Liu, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu. Efficient and effective regularized incomplete multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2634–2646, 2020. 6

[14] Xinwang Liu, Sihang Zhou, Yueqing Wang, Miaomiao Li, Yong Dou, En Zhu, and Jianping Yin. Optimal neighborhood kernel clustering with multiple kernels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 2

[15] Xinwang Liu, Xinzhong Zhu, Miaomiao Li, Lei Wang, En Zhu, Tongliang Liu, Marius Kloft, Dinggang Shen, Jianping Yin, and Wen Gao. Multiple kernel k k-means with incomplete kernels. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1191–1204, 2019. 1, 2

[16] Yue Liu, Xihong Yang, Sihang Zhou, Xinwang Liu, Zhen Wang, Ke Liang, Wenxuan Tu, Liang Li, Jingcan Duan, and Cancan Chen. Hard sample aware network for contrastive deep graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, 2023. 1

[17] Muhammad Mateen, Junhao Wen, Sun Song, and Zhouping Huang. Fundus image classification using vgg-19 architecture with pca and svd. *Symmetry*, 11(1):1, 2018. 1

[18] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001. 6

[19] Feiping Nie, Lai Tian, and Xuelong Li. Multiview clustering via adaptively weighted procrustes. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2022–2030, 2018. 6

[20] Erlin Pan and Zhao Kang. Multi-view contrastive graph clustering. *Advances in neural information processing systems*, 34:2148–2159, 2021. 1

[21] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *International conference on machine learning*, pages 5092–5101. PMLR, 2019. 1

[22] Jingkuan Song, Hanwang Zhang, Xiangpeng Li, Lianli Gao, Meng Wang, and Richang Hong. Self-supervised video hashing with hierarchical binary auto-encoder. *IEEE Transactions on Image Processing*, 27(7):3210–3221, 2018. 3

[23] Huayi Tang and Yong Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International Conference on Machine Learning*, pages 21090–21110. PMLR, 2022. 6

[24] Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1255–1265, 2021. 1

[25] Wenxuan Tu, Sihang Zhou, Xinwang Liu, Xifeng Guo, Zhiping Cai, En Zhu, and Jieren Cheng. Deep fusion clustering network. In *Proceedings of The Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 9978–9987, 2021. 1

[26] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *stat*, 1050:20, 2017. 1

[27] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Partial multi-view clustering via consistent gan. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1290–1295. IEEE, 2018. 3

[28] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Generative partial multi-view clustering. *arXiv preprint arXiv:2003.13088*, 2020. 3

[29] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30:1771–1783, 2021. 3

[30] Qianqian Wang, Huanhuan Lian, Gan Sun, Quanxue Gao, and Licheng Jiao. Icmsc: Incomplete cross-modal subspace clustering. *IEEE Transactions on Image Processing*, 30:305–317, 2020. 1, 3

[31] Siwei Wang, Xinwang Liu, Li Liu, Wenxuan Tu, Xinzhong Zhu, Jiyuan Liu, Sihang Zhou, and En Zhu. Highly-efficient incomplete large-scale multi-view clustering with consensus bipartite graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9776–9785, 2022. 1

[32] Siwei Wang, Xinwang Liu, Suyuan Liu, Jiaqi Jin, Wenxuan Tu, Xinzhong Zhu, and En Zhu. Align then fusion: Generalized large-scale multi-view clustering with anchor matching correspondences. *Advances in Neural Information Processing Systems*, 2022. 2

[33] Yiming Wang, Dongxia Chang, Zhiqiang Fu, Jie Wen, and Yao Zhao. Graph contrastive partial multi-view clustering. *IEEE Transactions on Multimedia*, 2022. 3

[34] Yiming Wang, Dongxia Chang, Zhiqiang Fu, Jie Wen, and Yao Zhao. Incomplete multi-view clustering via cross-view relation transfer. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 3

[35] Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*, 2016. 1

[36] Jie Wen, Zhihao Wu, Zheng Zhang, Lunke Fei, Bob Zhang, and Yong Xu. Structural deep incomplete multi-view clustering network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3538–3542, 2021. 3

[37] Jie Wen, Zheng Zhang, and Yong Xu. Cdimc-net: Cognitive deep incomplete multi-view clustering network. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3230–3236. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track. 1

[38] Jie Wen, Zheng Zhang, Yong Xu, Bob Zhang, Lunke Fei, and Hong Liu. Unified embedding alignment with missing views inferring for incomplete multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5393–5400, 2019. 1

[39] Jie Wen, Zheng Zhang, Zhao Zhang, Zhihao Wu, Lunke Fei, Yong Xu, and Bob Zhang. Dimc-net: Deep incomplete multi-view clustering network. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3753–3761, 2020. 1

[40] Wei Xia, Quanxue Gao, Qianqian Wang, Xinbo Gao, Chris Ding, and Dacheng Tao. Tensorized bipartite graph learning for multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[41] Xing Xu, Huimin Lu, Jingkuan Song, Yang Yang, Heng Tao Shen, and Xuelong Li. Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. *IEEE transactions on cybernetics*, 50(6):2400–2413, 2019. 3

[42] Zhe Xue, Junping Du, Changwei Zheng, Jie Song, Wenqi Ren, and Meiyu Liang. Clustering-induced adaptive struc-

ture enhancing network for incomplete multi-view data. In *IJCAI*, pages 3235–3241, 2021. 1

[43] Xihong Yang, Yue Liu, Sihang Zhou, Siwei Wang, Wenxuan Tu, Qun Zheng, Xinwang Liu, Liming Fang, and En Zhu. Cluster-guided contrastive graph clustering network. *arXiv preprint arXiv:2301.01098*, 2023. 1

[44] Yan Yang and Hao Wang. Multi-view clustering: A survey. *Big Data Mining and Analytics*, 1(2):83–107, 2018. 1

[45] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 3

[46] Changqing Zhang, Zongbo Han, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu, et al. Cpm-nets: Cross partial multi-view networks. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[47] Junpu Zhang, Liang Li, Siwei Wang, Jiyuan Liu, Yue Liu, Xinwang Liu, and En Zhu. Multiple kernel clustering with dual noise minimization. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3440–3450, New York, NY, USA, 2022. Association for Computing Machinery. 1

[48] Pei Zhang, Siwei Wang, Jingtao Hu, Zhen Cheng, Xifeng Guo, En Zhu, and Zhiping Cai. Adaptive weighted graph fusion incomplete multi-view subspace clustering. *Sensors*, 20(20):5755, 2020. 2

[49] Handong Zhao, Hongfu Liu, and Yun Fu. Incomplete multimodal visual data grouping. In *IJCAI*, pages 2392–2398, 2016. 2

[50] Pengfei Zhu, Binyuan Hui, Changqing Zhang, Dawei Du, Longyin Wen, and Qinghua Hu. Multi-view deep subspace clustering networks. *arXiv preprint arXiv:1908.01978*, 2019. 3