COMPLETER: Incomplete Multi-view Clustering via Contrastive Prediction

Yijie Lin[1], Yuanbiao Gou[1], Zitao Liu[2], Boyun Li[1], Jiancheng Lv[1], Xi Peng[1*]
[1] College of Computer Science, Sichuan University, China.
[2] TAL Education Group, China.
{linyijie.gm,gouyuanbiao,zitao.jerry.liu,liboyun.gm,pengx.gm}@gmail.com; lvjiancheng@scu.edu.cn

Abstract

In this paper, we study two challenging problems in incomplete multi-view clustering analysis, namely, i) how to learn an informative and consistent representation among different views without the help of labels and ii) how to recover the missing views from data. To this end, we propose a novel objective that incorporates representation learning and data recovery into a unified framework from the view of information theory. To be specific, the informative and consistent representation is learned by maximizing the mutual information across different views through contrastive learning, and the missing views are recovered by minimizing the conditional entropy of different views through dual prediction. To the best of our knowledge, this could be the first work to provide a theoretical framework that unifies the consistent representation learning and cross-view data recovery. Extensive experimental results show the proposed method remarkably outperforms 10 competitive multi-view clustering methods on four challenging datasets. The code is available at https://pengxi.me.

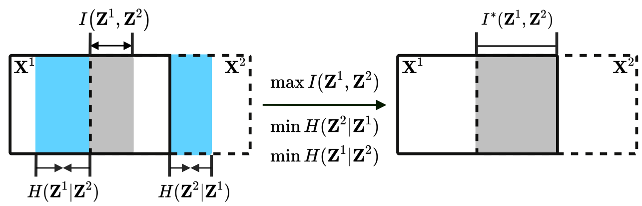


Figure 1. Our basic observation and theoretical results from the perspective of information theory. In the figure, the solid and dotted rectangles denote the information contained in view 1 ($\mathbf{X}^1$) and view 2 ($\mathbf{X}^2$), respectively. In mathematical, the mutual information $I(\mathbf{Z}^1, \mathbf{Z}^2)$ (grey area) quantifies the amount of information shared by $\mathbf{Z}^1$ and $\mathbf{Z}^2$, where $\mathbf{Z}^1$ and $\mathbf{Z}^2$ are the representations of $\mathbf{X}^1$ and $\mathbf{X}^2$, respectively. To learn consistent representations, it is encouraged to maximize $I(\mathbf{Z}^1, \mathbf{Z}^2)$. In addition, minimizing the conditional entropy $H(\mathbf{Z}^i|\mathbf{Z}^j)$ (blue area) will encourage the recovery of missing view because $\mathbf{Z}^i$ is fully determined by $\mathbf{Z}^j$ if and only if the conditional entropy $H(\mathbf{Z}^i|\mathbf{Z}^j) = 0$, where $i = 1$, $j = 2$ or $i = 2$, $j = 1$. Subtly, on the one hand, the maximization of $I(\mathbf{Z}^1, \mathbf{Z}^2)$ could increase the amount of the shared information, thus the data recoverability could be benefited, *i.e.*, it is easier to recover one view from the other. On the other hand, as $H(\mathbf{Z}^i|\mathbf{Z}^j)$ quantifies the amount of information of $\mathbf{Z}^i$ conditioned on $\mathbf{Z}^j$, the minimization of $H(\mathbf{Z}^i|\mathbf{Z}^j)$ will encourage to discard the inconsistent information across-views, and thus the consistency could be further improved. With the above observation, cross-view consistency and data recovery are treated as two sides of one coin under the above unified information theory framework.

1. Introduction

In the real world, multi-view data, which often exhibit heterogeneous properties, is collected from diverse sensors or obtained from various feature extractors. As one of the most important unsupervised multi-view methods, multi-view clustering (MVC) aims to separate data points into different clusters in an unsupervised fashion [11, 17, 20, 29, 40, 54]. To achieve the end, the key is exploring the consistency across different views so that a common/shared representation is learned [5, 12, 14, 21, 33, 47]. Behind the consistency learning, the implicit assumption is that the views are complete, *i.e.*, all data points will present in all possible views.

In practice, however, some views of data points might be missing due to the complexity in data collection and transmission, leading to so-called *incomplete multi-view problem (IMP)*. For example, in online meetings, some video frames might lose the visual or audio signal due to the breakdown of sensors. To solve IMP, some incomplete multi-view clustering algorithms (IMC) have been proposed by employing numerous data recovery methods to complete the missing data, *e.g.*, matrix factorization based methods [10, 22, 35, 46, 53] and generative adversarial networks based methods [16, 41, 45]. These works have attempted to overcome the following two challenges: i) how to learn informative and consistent representations across different views? and ii) how to eliminate the influence of the miss-

*Corresponding author

ing views? Although some promising results have been achieved, almost all existing works treat these two challenges as two independent problems and a unified theoretical understanding is still lacking.

Different from existing IMC studies, we theoretically show that cross-view consistency learning and data recovery could be treated as two sides of one coin and these two challenging tasks could mutually boost. Our motivation comes from [38], as shown in Fig. 1. It should be pointed out that, [38] utilizes predictive learning to enhance the performance of contrastive learning, while we aim at recovering the missing data through dual prediction. Moreover, another difference lies on our theoretical result, *i.e.*, the data recovery and consistency learning could mutually boost through contrastive learning and dual prediction.

Based on our observations and theoretical results, we propose a novel incomplete multi-view clustering method, termed inCOMPlete muLti-view clustEring via conTrastivE pRediction (COMPLETER). In detail, COMPLETER projects a given dataset into a feature space wherein information consistency and data restorability are guaranteed using three jointly learning objectives. More specifically, a within-view reconstruction loss is used to learn a view-specific representation so that the trivial solution is avoided. In the latent feature space, a contrastive loss is introduced to learn the cross-view consistency by maximizing mutual information $I(\mathbf{Z}^1, \mathbf{Z}^2)$, and a dual prediction loss is used to recover the missing view by minimizing conditional entropy $H(\mathbf{Z}^1|\mathbf{Z}^2)$ and $H(\mathbf{Z}^2|\mathbf{Z}^1)$. It should be pointed out that the data recovery referred in this paper is task-oriented, *i.e.*, only the shared instead of all information would be recovered to facilitate the downstream tasks like MVC. To summarize:

- We provide a novel insight to the community, *i.e.*, the data recovery and consistency learning of incomplete multi-view clustering are with intrinsic connections, which could be elegantly unified into the framework of information theory. Such a theoretical view is remarkably different from existing works which treat consistency learning and data recovery as two separate problems.

- The proposed COMPLETER method is with a novel loss function which achieves the information consistency and data restorability using a contrastive loss and a dual prediction loss. Extensive experiments verify the effectiveness of the proposed loss function.

2. Related Work

In this section, we briefly review some recent developments in two related topics, namely, incomplete multi-view clustering and contrastive learning.

2.1. Incomplete Multi-view Clustering

Based on the way of utilizing the multi-view information, most existing IMC methods could be roughly classified into three categories, *i.e.*, matrix factorization (MF) based IMC [10,22,35,53], spectral clustering based IMC [39], and kernel learning based IMC [26]. In brief, MF based methods project the incomplete data into a common subspace by utilizing the low-rankness. For example, DAIMC [10] establishes a consensus basis matrix with the help of $\ell_{2,1}$-norm and IMG [53] utilizes the ℓ_F-norm to reduce the influence of missing data. As a typical spectral clustering based method, PIC [39] learns the common representation using a consistent Laplacian graph constructed from incomplete views. EERIMVC [26] proposes using a multi-kernel method to achieve IMC in an iterative optimization manner. Besides, the methods like [16, 41] utilize cycleGAN [55] to generate the missing view from the complete views and CDIMC-net [44] incorporates the view-specific encoders and the graph embedding strategy to handle the incomplete multi-view data.

The differences between this study and existing works are given below. First, we aim to infer the missing data rather than the missing similarity, thus enjoying higher interpretability [26]. Second, our method is a deep rather than shallow model [10, 19, 22, 26, 35, 39, 53], thus naturally embracing the capacity of handling complex and large-scale dataset. Third, almost all existing IMC methods [10,16,22,26,35,39,41,53] treat data recovery and consistency learning as two independent problems/steps, while lacking a theoretical understanding. In contrast, we proposed that data recovery and consistency learning could be unified into the framework of information theory [36]. Both data recovery and consistency learning could be of benefit to learning the common representation.

2.2. Contrastive Learning

As one of most effective unsupervised learning paradigms, contrastive learning [2, 4, 8, 23, 28, 30, 37, 38] has achieved state-of-the-art performance in representation learning. The basic idea of contrastive learning is learning a feature space from raw data by maximizing the similarity between positive pairs while minimizing that of negative pairs. In recent, some studies show that the success of contrastive learning could attribute to the maximization of mutual information. For example, MoCo [9] and CPC [30] minimize the InfoNCE loss that can be regarded as maximizing a lower bound on mutual information, *i.e.*, $I\left(\mathbf{Z}^1, \mathbf{Z}^2\right) \geq \log(N) - \mathcal{L}_{\mathrm{NCE}}$, where N is the number of negative pairs, $\mathbf{Z}^1$ and $\mathbf{Z}^2$ are the latent representations of multi-view data $\mathbf{X}^1$ and $\mathbf{X}^2$, respectively.

The differences between this work and existing contrastive learning studies are as below. First, most existing contrastive learning methods [2, 8, 9, 28] aim to handle

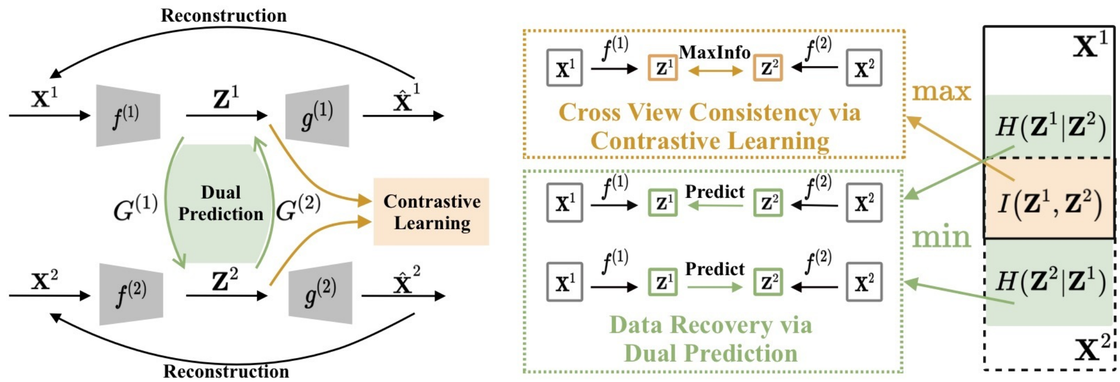


Figure 2. Overview of COMPLETER. In the figure, bi-view data is used as a showcase. As shown, our method contains three joint learning objectives, *i.e.*, within-view reconstruction, cross-view contrastive learning, and cross-view dual prediction. To be specific, the within-view reconstruction objective aims to project all views into view-specific spaces with the minimal reconstruction loss. The cross-view contrastive learning objective is implemented by maximizing the mutual information between $\mathbf{Z}^1$ and $\mathbf{Z}^2$. The cross-view dual prediction objective utilizes two mapping $G^{(1)}$ and $G^{(2)}$ to recover one view from another one by minimizing the conditional entropy $H(\mathbf{Z}^i|\mathbf{Z}^j)$.

single-view data and exhaustively explore a variety of data augmentations to build different views/augmentations. In contrast, our method aims to learn consistency from a given multi-view dataset. To the best of our knowledge, this could be one of the first studies on multi-view contrastive learning. Second, our method is specifically designed for handling missing data, whereas the existing contrastive learning works ignore this practical problem. Third, although existing contrastive learning studies have shown that the consistency could be learned by maximizing the mutual information of different augmentations, they ignore the inconsistency learning. With a unified framework of information theory, we show that inconsistency learning could be defined by conditional entropy and the missing data could be recovered through the minimization of the inconsistency.

3. The Proposed Method

In this section, we propose a deep multi-view clustering method, termed inCOMPlete muLti-view clustEring via conTrastivE pRediction (COMPLETER) for learning the representations with a set of incomplete multi-view samples. As illustrated in Fig. 2, COMPLETER consists of three jointly learning objectives, namely, within-view reconstruction, cross-view contrastive learning, and cross-view dual prediction. For clarity, we will first introduce the proposed loss function and then elaborate on each objective.

3.1. The Objective Function

Without loss of generality, we take bi-view data as an example. Given a dataset $\bar{\mathbf{X}} = \{\bar{\mathbf{X}}^{1,2}, \bar{\mathbf{X}}^1, \bar{\mathbf{X}}^2\}$ of n instances, where $\bar{\mathbf{X}}^{1,2}$, $\bar{\mathbf{X}}^1$, and $\bar{\mathbf{X}}^2$ denote the examples presented in both views, the first view only, and the second view only, respectively. Let m be the data size of complete examples $\bar{\mathbf{X}}^{1,2}$ and $\mathbf{X}^v$ be the v-th view of $\bar{\mathbf{X}}^{1,2}$, then $\bar{\mathbf{X}}^{1,2} = \{\mathbf{X}^1, \mathbf{X}^2\}$.

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

$$\mathcal{L} = \mathcal{L}_{cl} + \lambda_1 \mathcal{L}_{pre} + \lambda_2 \mathcal{L}_{rec}, \tag{1}$$

where $\mathcal{L}_{cl}$, $\mathcal{L}_{pre}$, and $\mathcal{L}_{rec}$ are cross-view contrastive loss, dual prediction loss, and within-view reconstruction loss, respectively. The parameters λ_1 and λ_2 are the balanced factors on $\mathcal{L}_{pre}$, and $\mathcal{L}_{rec}$, respectively. In our experiments, we simply fix these two parameters to 0.1.

Within-view Reconstruction: For each view, we pass it through an autoencoder to learn the latent representation $\mathbf{Z}^v$ by minimizing

$$\mathcal{L}_{rec} = \sum_{v=1}^{2} \sum_{t=1}^{m} \left\| \mathbf{X}_t^v - g^{(v)} \left(f^{(v)} \left(\mathbf{X}_t^v \right) \right) \right\|_2^2, \tag{2}$$

where $\mathbf{X}_t^v$ denotes the t-th sample of $\mathbf{X}^v$. $f^{(v)}$ and $g^{(v)}$ denote the encoder and decoder for the v-th view, respectively. Hence, the representation of t-th sample in v-th view is given by

$$\mathbf{Z}_t^v = f^{(v)} \left(\mathbf{X}_t^v \right), \tag{3}$$

where $\mathbf{Z}^v$ denotes the representations of $\mathbf{X}^v$ and $v \in \{1, 2\}$.

It should be pointed out that the autoencoder structure is helpful to avoid the trivial solution.

Cross-view Contrastive Learning: In the latent space parameterized by $\mathcal{L}_{rec}$, we conduct contrastive learning to learn a common representation shared across different views. Unlike most existing contrastive learning studies [9, 30] which maximize the consistency between the learned representations $\mathbf{Z}^1$ and $\mathbf{Z}^2$ by maximizing the lower bound of mutual information, we directly maximize the mutual information between the representations of different views. Mathematically,

$$\mathcal{L}_{cl} = -\sum_{t=1}^{m} \left(I\left(\mathbf{Z}_t^1, \mathbf{Z}_t^2\right) + \alpha\left(H\left(\mathbf{Z}_t^1\right) + H\left(\mathbf{Z}_t^2\right)\right)\right), \quad (4)$$

where I denotes the mutual information, H is the information entropy, and parameter α is set as 9 to regularize the entropy in our experiments. We design this objective with the following goals. On the one hand, from information theory, information entropy is the average amount of information conveyed by an event [3]. Hence a larger entropy $H(\mathbf{Z}^i)$ denotes a more informative representation $\mathbf{Z}^i$. On the other hand, the maximization of $H(\mathbf{Z}^1)$ and $H(\mathbf{Z}^2)$ will avoid the trivial solution of assigning all samples to the same cluster.

To formulate $I(\mathbf{Z}_t^1, \mathbf{Z}_t^2)$, we first define the joint probability distribution $\mathcal{P}(z, z')$ of variable z and z'. As a softmax function is stacked at the last layer of the encoder, each element of $\mathbf{Z}^1$ and $\mathbf{Z}^2$ could be regarded as an over-cluster class probability like [13,15,34]. In other words, $\mathbf{Z}^1$ and $\mathbf{Z}^2$ could be understood as the distribution of two discrete cluster assignment variables z and z' over D "classes", where D is the dimension of $\mathbf{Z}^1$ and $\mathbf{Z}^2$. As a result, $\mathcal{P}(z, z')$ is defined as $\mathbf{P} \in \mathcal{R}^{D\times D}$, *i.e.*,

$$\mathbf{P} = \frac{1}{m}\sum_{t=1}^{m} \mathbf{Z}_t^1 \left(\mathbf{Z}_t^2\right)^\top. \quad (5)$$

Let $\mathbf{P}_d$ and $\mathbf{P}'_d$ denote the marginal probability distributions $\mathcal{P}(z = d)$ and $\mathcal{P}(z' = d')$, they could be obtained by summing over the d-th rows and d'-th columns of joint probability distribution matrix $\mathbf{P}$. Expecting z and z' are with equal importance, $\mathbf{P}$ is further calculated by $(\mathbf{P} + \mathbf{P}^T)/2$. For discrete distributions, Eq. (4) is given as below:

$$\mathcal{L}_{cl} = -\sum_{d=1}^{D}\sum_{d'=1}^{D} \mathbf{P}_{dd'} \ln \frac{\mathbf{P}_{dd'}}{\mathbf{P}_d^{\alpha+1}\cdot \mathbf{P}_{d'}^{\alpha+1}}, \quad (6)$$

where $\mathbf{P}_{dd'}$ is the element at the d-th row and d'-th column of $\mathbf{P}$ and α is a balance parameter of entropy as defined in Eq. (4). The details from Eq. (4) to Eq. (6) are presented in supplementary material.

Cross-view Dual Prediction: To infer the missing views, we propose a dual prediction mechanism as shown in Fig. 2. To be specific, in a latent space parameterized by a neural network, the view-specific representation will be predicted by another through minimizing the entropy $H(\mathbf{Z}^i|\mathbf{Z}^j)$, where $i = 1$, $j = 2$ or $i = 2$, $j = 1$. Such a dual prediction mechanism is with theoretical explanation as elaborated in Fig. 1. In short, $\mathbf{Z}^i$ is fully determined by $\mathbf{Z}^j$ if and only if the conditional entropy $H(\mathbf{Z}^i|\mathbf{Z}^j) = -\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}[\log \mathcal{P}(\mathbf{Z}^i|\mathbf{Z}^j)] = 0$. To solve this objective, a common approximative approach is introducing a variational distribution $\mathcal{Q}(\mathbf{Z}^i|\mathbf{Z}^j)$ and maximizing $\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}\left[\log \mathcal{Q}\left(\mathbf{Z}^i \mid \mathbf{Z}^j\right)\right]$ which is the lower bound of $\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}[\log \mathcal{P}(\mathbf{Z}^i|\mathbf{Z}^j)]$, *i.e.*,

$$\begin{aligned}\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}\left[\log \mathcal{P}\left(\mathbf{Z}^i \mid \mathbf{Z}^j\right)\right] =& \mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}\left[\log \mathcal{Q}\left(\mathbf{Z}^i \mid \mathbf{Z}^j\right)\right] + \\ & D_{\mathrm{KL}}\left(\mathcal{P}\left(\mathbf{Z}^i \mid \mathbf{Z}^j\right) \| \mathcal{Q}\left(\mathbf{Z}^i \mid \mathbf{Z}^j\right)\right).\end{aligned} \quad (7)$$

Such a variational distribution $\mathcal{Q}$ can be any types like Gaussian [7] and Laplacian distribution [55]. In practice, we simply assume the distribution $\mathcal{Q}$ as a Gaussian distribution $\mathcal{N}\left(\mathbf{Z}^i \mid G^{(j)}\left(\mathbf{Z}^j\right), \sigma\mathbf{I}\right)$, where $G^{(j)}(\cdot)$ could be a parameterized model which maps $\mathbf{Z}^j$ to $\mathbf{Z}^i$ and $\sigma\mathbf{I}$ is the variance matrix. By ignoring the constants derived from the Gaussian distribution, maximizing $\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}\left[\log \mathcal{Q}\left(\mathbf{Z}^i \mid \mathbf{Z}^j\right)\right]$ is equivalent to

$$\min \mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}} \left\|\mathbf{Z}^i - G^{(j)}\left(\mathbf{Z}^j\right)\right\|_2^2. \quad (8)$$

For a given bi-view dataset, we further have

$$\mathcal{L}_{pre} = \left\|G^{(1)}(\mathbf{Z}^1) - \mathbf{Z}^2\right\|_2^2 + \left\|G^{(2)}(\mathbf{Z}^2) - \mathbf{Z}^1\right\|_2^2. \quad (9)$$

It should be pointed out that the above loss may lead to trivial solutions without the within-view reconstruction loss, *i.e.*, $\mathbf{Z}^1$ and $\mathbf{Z}^2$ are equivalent to the same constant.

After the model converged, it is easy to predict the missing representation $\bar{\mathbf{Z}}^i$ from $\bar{\mathbf{Z}}^j$ through the above dual mapping, *i.e.*,

$$\bar{\mathbf{Z}}^i = G^{(j)}\left(\bar{\mathbf{Z}}^j\right) = G^{(j)}\left(f^{(j)}(\bar{\mathbf{X}}^j)\right), \quad (10)$$

where $\bar{\mathbf{Z}}^j$ denotes the representations of $\bar{\mathbf{X}}^j$.

3.2. Implementation Details

As shown in Fig. 2, COMPLETER consists of two training modules, *i.e.*, two view-specific autoencoders and two cross-view prediction networks. For these two modules, we simply adopt a fully-connected network where each layer is followed by a batch normalization layer and a ReLU layer. The softmax activation function is used at the last layer of the encoder and prediction module. In the supplementary material, all details of our model have been presented.

In the training stage, we use the complete data $\bar{\mathbf{X}}^{1,2}$ to train COMPLETER in an end-to-end manner. Specifically, we train the autoencoders by $\mathcal{L}_{cl}$ and $\mathcal{L}_{rec}$ in the first 100 epochs to stabilize the training of the dual prediction. Then, we train the whole networks with $\mathcal{L}$ for 400 epochs. Once the network converged, we feed the whole dataset into the network to obtain the representations for all views including

Table 1. The clustering performance comparisons on four challenging datasets. "–" indicates unavailable results due to out of memory. The 1st/2nd best results are indicated in red/blue.

Missing Type	Method	Caltech101-20			LandUse-21			Scene-15			Noisy MNIST		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Incomplete	AE^{2}Nets [51]	33.61	49.20	24.99	19.22	23.03	5.75	27.88	31.35	13.93	38.67	33.79	19.99
	IMG [53]	42.29	58.26	33.69	15.52	22.54	3.73	23.96	25.70	9.21	–	–	–
	UEAF [43]	47.35	56.71	37.08	16.38	18.42	3.80	28.20	27.01	8.70	34.56	33.13	24.04
	DAIMC [10]	44.63	59.53	32.70	19.30	19.45	5.80	23.60	21.88	9.44	34.44	27.15	16.42
	EERIMVC [26]	40.66	51.38	27.91	22.14	25.18	9.10	33.10	32.11	15.91	54.97	44.91	35.94
	DCCAE [42]	40.01	52.88	30.00	14.94	20.94	3.67	31.75	34.42	15.80	61.79	59.49	33.49
	PVC [22]	41.42	56.53	31.00	21.33	23.14	8.10	25.61	25.31	11.25	35.97	27.74	16.99
	BMVC [52]	32.13	40.58	12.20	18.76	18.73	3.70	30.91	30.23	10.93	24.36	15.11	6.50
	DCCA [1]	38.59	52.51	29.81	14.08	20.02	3.38	31.83	33.19	14.93	61.82	60.55	37.71
	PIC [39]	57.53	64.32	45.22	23.60	26.52	9.45	38.70	37.98	21.16	–	–	–
	COMPLETER	68.44	67.39	75.44	22.16	27.00	10.39	39.50	42.35	23.51	80.01	75.23	70.66
Complete	AE^{2}Nets [51]	49.10	65.38	35.66	24.79	30.36	10.35	36.10	40.39	22.08	56.98	46.83	36.98
	IMG [53]	44.51	61.35	35.74	16.40	27.11	5.10	24.20	25.64	9.57	–	–	–
	UEAF [43]	47.40	57.90	38.98	23.00	27.05	8.79	34.37	36.69	18.52	67.33	65.37	55.81
	DAIMC [10]	45.48	61.79	32.40	24.35	29.35	10.26	32.09	33.55	17.42	39.18	35.69	23.65
	EERIMVC [26]	43.28	55.04	30.42	24.92	29.57	12.24	39.60	38.99	22.06	65.47	57.69	49.54
	DCCAE [42]	44.05	59.12	34.56	15.62	24.41	4.42	36.44	39.78	21.47	81.60	84.69	70.87
	PVC [22]	44.91	62.13	35.77	25.22	30.45	11.72	30.83	31.05	14.98	41.94	33.90	22.93
	BMVC [52]	42.55	63.63	32.33	25.34	28.56	11.39	40.50	41.20	24.11	81.27	76.12	71.55
	DCCA [1]	41.89	59.14	33.39	15.51	23.15	4.43	36.18	38.92	20.87	85.53	89.44	81.87
	PIC [39]	62.27	67.93	51.56	24.86	29.74	10.48	38.72	40.46	22.12	–	–	–
	COMPLETER	70.18	68.06	77.88	25.63	31.73	13.05	41.07	44.68	24.78	89.08	88.86	85.47

the missing ones. After that, the common representation, which is obtained by simply concatenating all view-specific representations together, is further fed into k-means to get the clustering results like the traditional fashion [1, 10, 22, 25, 26, 32, 39, 42, 43, 48, 49, 53].

4. Experiments

In this section, we evaluate the proposed COMPLETER method on four widely-used multi-view datasets with the comparisons of 10 multi-view clustering approaches.

4.1. Experimental Settings

Four widely-used datasets are used in our experiments. In brief, **Caltech101-20** [24] consists of 2,386 images of 20 subjects with the views of HOG and GIST features. **Scene-15** [6], which consists of 4,485 images distributed over 15 scene categories, is with PHOG and GIST features. **LandUse-21** [50] consists of 2100 satellite images from 21 categories with PHOG and LBP features. **Noisy MNIST** [42] uses the original 70k MNIST images as view 1 and randomly selects within-class images with white Gaussian noise as view 2. As most of the baselines cannot handle such a large dataset, we could only use a subset of Noisy MNIST consisting of 10k validation images and 10k testing images.

To evaluate the performance of handling incomplete multi-view data, we randomly select some instances as incomplete data by randomly removing one view. The missing rate η is defined as $\eta = (n - m)/n$, where m is the number of complete examples, and n is the number of the whole dataset.

For a comprehensive analysis, three widely-used clustering metrics including Normalized Mutual Information (NMI), Accuracy (ACC), and Adjusted Rand Index (ARI) are used. A higher value of these metrics indicates a better clustering performance.

We implement our COMPLETER in PyTorch 1.2 [31] and carry all evaluations on a standard Ubuntu-18.04 OS with an NVIDIA 2080Ti GPU. We use Adam optimizer [18] with the default parameters to train our model and set the initial learning rate as 0.0001. The batch size is set to 256 and the maximal training epoch is fixed to 500 on all datasets. The entropy parameter α is fixed to 9 and trade-off hyper-parameters λ_1 and λ_2 are fixed to 0.1 for all datasets. In our implementation environment, COMPLETER takes about 60 seconds to train a model on Caltech101-20, 80 seconds on Scene-15, 50 seconds on LandUse-15, and 500 seconds on NoisyMNIST.

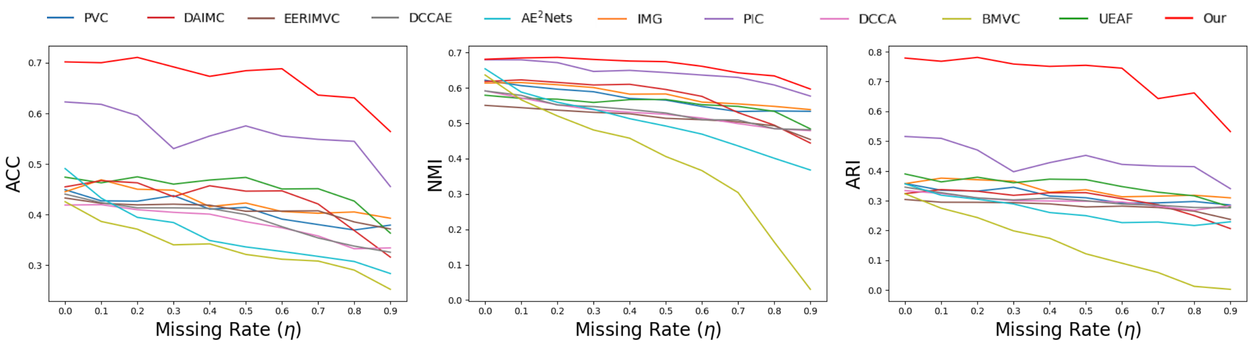


Figure 3. Performance comparisons on Caltech101-20 with different missing rates (η).

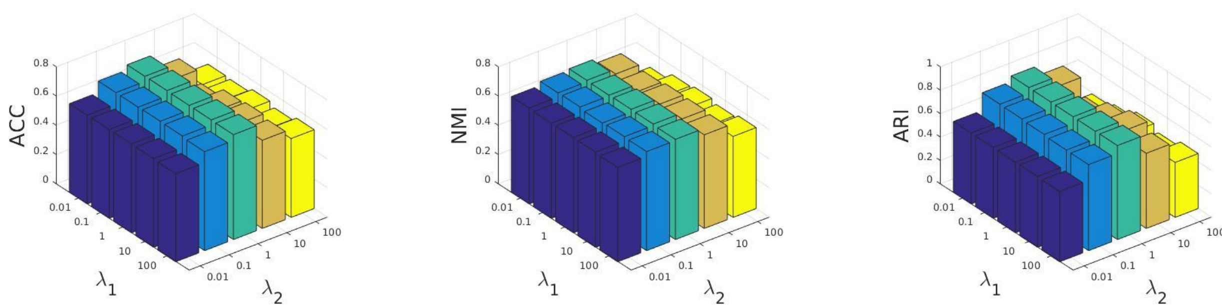


Figure 4. Parameter analysis on Caltech101-20.

4.2. Comparisons with State of the Arts

We compare COMPLETER with 10 multi-view clustering approaches including Deep Canonically Correlated Analysis (DCCA) [1], Deep Canonically Correlated Autoencoders (DCCAE) [42], Binary Multi-view Clustering (BMVC) [52], Autoencoder in Autoencoder Networks (AE^2-Nets) [51], Partial Multi-View Clustering (PVC) [22], Efficient and Effective Regularized Incomplete Multi-view Clustering (EERIMVC) [26], Doubly Aligned Incomplete Multi-view Clustering (DAIMC) [10], Incomplete Multi-Modal Visual Data Grouping (IMG) [53], Unified Embedding Alignment Framework (UEAF) [43], and Perturbation-oriented Incomplete Multi-view Clustering (PIC) [39]. The first four methods could only handle complete multi-view data and thus we fill the missing data with the mean values of the same view. For all methods, we use the recommended network structure and parameters for fair comparisons. In brief, for CCA-based methods (*i.e.*, DCCA and DCCAE), we fix the hidden representation dimension to 10. For BMVC, we fix the length of binary code to 128. For EERIMVC, we exploit the “Gauss kernel” to construct the kernel matrices and seek the optimal λ from 2^{-15} to 2^{15} with an interval of 2^3.

We test all methods in two settings, *i.e.*, missing rate $\eta = 0.5$ (denoted by *Incomplete*) and $\eta = 0$ (denoted by *Complete*). The average clustering results are obtained by repeating each method with five random initializations and dataset partitions.

As shown in Table 1, COMPLETER significantly outperforms these state-of-the-art baselines by a large performance margin on all four datasets. In the *Incomplete* setting, COMPLETER surpasses the best baseline by 3.07% on Caltech101-20, 4.37% on Scene-15, and 14.68% on NoisyMNIST in terms of NMI. Moreover, COMPLETER achieves more than 50% performance improvements over the best baseline on Caltech101-20 and NoisyMNIST in terms of ARI. In the *Complete* setting, COMPLETER also remarkably outperforms almost all baselines. The encouraging performance demonstrates the promising representability of COMPLETER thanks to our unified theoretical framework of contrastive learning and dual prediction.

4.3. Performance with Different Missing Rates

To further investigate the effectiveness of our method, we conduct experiments by varying the missing rate η from 0 to 0.9 with a gap of 0.1 on Caltech101-20. When the missing rate is 0.9, the size of the whole training data is smaller than that of a data batch, and thus we reduce the batch size to 128. From the results in Fig. 3, one could observe that: i) COMPLETER significantly outperforms all

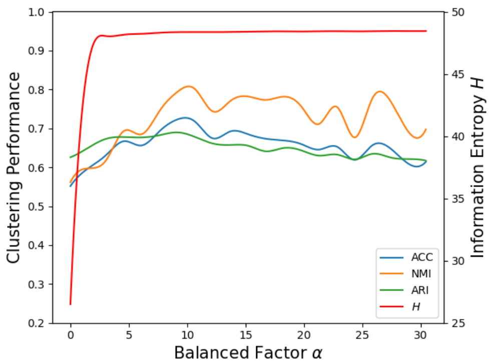


Figure 5. Clustering results of COMPLETER with the increase of entropy α on Caltech101-20. The x-axis denotes α, the left and right y-axis denote clustering performance and information entropy, respectively.

the tested baselines in all missing rates setting; ii) with increasing the missing rate, the performance degradations of the compared methods are much larger than that of ours. For example, COMPLETER and PIC achieve the NMI of 0.6806 and 0.6793 with $\eta = 0$, respectively, while with the increase of the missing rate, COMPLETER is remarkably superior to PIC.

4.4. Parameter Analysis and Ablation Studies

In this section, we analyze COMPLETER on the Caltech101-20 dataset from two perspectives, *i.e.*, parameter sensitivity analysis and ablation studies. In the evaluations, the missing rate η is fixed to 0.5.

Our method contains three user-specified parameters, *i.e.*, the entropy parameter α, the prediction trade-off parameter λ_1, and the reconstruction trade-off parameter λ_2. In the following studies, we first investigate the relation among α, information entropy of representations $H(Z^i)$, and clustering performance by fixing λ_1 and λ_2 to 0.1 and changing the value of α. As shown in Fig. 5, the information entropy grows in step with α. Specifically, with the increase of the information entropy (from left to right), the clustering performance (ACC, NMI, and ARI) improves first and then degrades. The reason may due to the following aspects. On the one hand, the increased entropy (information contained in the representation) will enlarge the mutual information which further boosts the clustering performance. On the other hand, with the increase of α, an over-informative representation will suppress the mutual information term in Eq. (4) and then the consistency is reduced.

To evaluate the influence of λ_1 and λ_2, we change their value in the range of $\{0.01, 0.1, 1, 10, 100\}$. As shown in Fig. 4, our method is robust to the choice of λ_1. In addition, a good choice of λ_2 will remarkably improve the performance of COMPLETER.

Table 2. Ablation study on Caltech101-20. In the table, "✓" denotes COMPLETER with the component.

$\mathcal{L}_{pre}$	$\mathcal{L}_{cl}$	$\mathcal{L}_{rec}$	ACC	NMI	ARI
		✓	33.65	31.60	16.43
✓			38.61	37.65	26.50
	✓		46.69	58.03	41.86
	✓	✓	55.75	59.35	58.88
✓		✓	54.70	52.63	43.49
✓	✓		64.59	62.11	71.07
✓	✓	✓	**68.44**	**67.39**	**75.44**

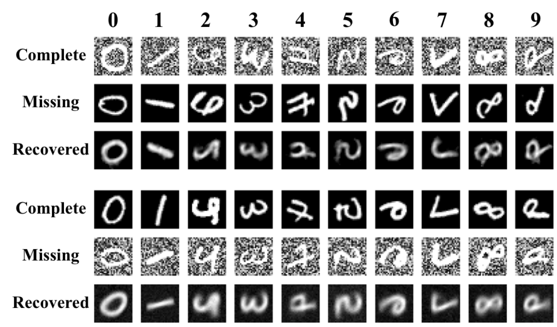


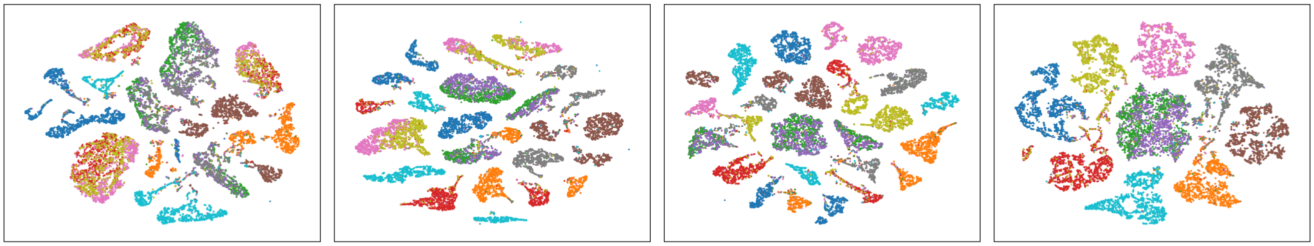
Figure 6. Data recovery on Noisy MNIST. Line 1 and 4 are complete views, Line 2 and 5 are missing views, and Line 3 and 6 are the recovered results from the complete view.

To further verify the importance of each module in COMPLETER, we conduct the following ablation study. In detail, the following seven experiments are designed to isolate the effect of the contrastive loss $\mathcal{L}_{cl}$, the reconstruction loss $\mathcal{L}_{rec}$, and the dual prediction loss $\mathcal{L}_{pre}$. As shown in Table 2, all loss terms play indispensable roles in COMPLETER. It should be pointed out that optimizing the dual prediction loss $\mathcal{L}_{pre}$ alone may lead to trivial solutions. To solve this problem, we add a batch normalization layer to each fully-connected layer and report the corresponding results.

4.5. Visualization Verification on Our Theoretical Results

In this section, we carry out experiments to verify our theoretical results presented in Fig. 1. The experiments are conducted on Noisy MNIST dataset by visualizing the recovered views and the common representations. In the experiments, the missing rate η is fixed to 0.5.

Different from most existing incomplete multi-view methods, COMPLETER could explicitly infer the representation of the missing views. As a result, the corresponding reconstruction in the original space could be obtained through the decoder. To show the recoverability of COMPLETER, Fig. 6 visually illustrates some recovered exam-



(a) 20 epoch (NMI = 0.567) (b) 50 epoch (NMI = 0.651) (c) 100 epoch (NMI = 0.707) (d) 200 epoch (NMI = 0.759)

Figure 7. t-sne visualization on the Noisy MNIST dataset with increasing training iteration.

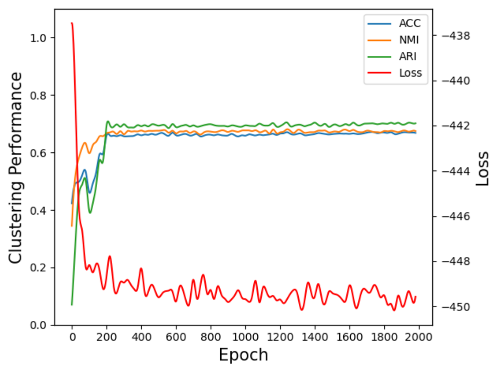


Figure 8. Clustering performance of COMPLETER with increasing epoch on Caltech101-20. The x-axis denotes the training epoch, the left and right y-axis denote the clustering performance and corresponding loss value, respectively.

ples from Noisy MNIST. From the results, one could have the following observations. In the top three rows, the recovered images (Line 3) are much similar to the complete ones (Line 1), while being with a clean background like the missing view (Line 2). In the bottom three rows, a similar observation could also be obtained even though COMPLETER recovered the missing images from the images with a clean rather than noisy background. In short, COMPLETER could recover the important information while discarding the indistinct characteristics like noises in this example.

It should be noticed that the semantic information and the noisy background in this example could be regarded as consistency and inconsistency of two views. Therefore, the reasons for the above observations are two-fold. On the one hand, the recovered views will contain the shared information (semantic information instead of noise) of two available views thanks to the maximization of the mutual information. On the other hand, the minimization of conditional entropy designed for data recovery could subtly discard the inconsistent information across different views. As a result, the noise in the missing views will be suppressed during recovery. This verifies the effectiveness of our theory.

Besides the above visualizations, we also demonstrate the t-sne [27] visualizations of the learned common representations. As shown in Fig. 7, the learned representations become more compact and discriminative with the increase of the epoch.

4.6. Convergence Analysis

In this section, we investigate the convergence of COMPLETER by reporting the loss value and the corresponding clustering performance with increasing epochs. As shown in Fig. 8, one could observe that the loss remarkably decreases in the first 200 epochs, and meanwhile ACC, NMI, and ARI continuously increase.

5. Conclusion

To learn common representations from a given multi-view data wherein some views are missing, this paper proposed COMPLETER which embraces the rigid mathematical motivation and explanation from information theory. In short, we treat consistency learning and view completing as two sides of one coin rather than two separate problems. Such a unified framework would provide novel insight to the community on understanding consistency learning and data recovery. In the future, we plan to further explore the potential of our theoretical framework in other multi-view learning tasks, *e.g.*, object ReId. Moreover, it is also promising to extend it to handle the image translation tasks.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2020AAA0104500; in part by the Fundamental Research Funds for the Central Universities under Grant YJ201949; in part by NFSC under Grant U19A2081, 61625204, 61836006, U19A2078; in part by the Fund of Sichuan University Tomorrow Advancing Life; and in part by Beijing Nova Program (Z201100006820068) from Beijing Municipal Science & Technology Commission.

References

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, pages 1247–1255, 2013. 5, 6

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv:2002.05709*, 2020. 2

[3] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. 2006. 4

[4] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018. 2

[5] Cheng Deng, Zongting Lv, Wei Liu, Junzhou Huang, Dacheng Tao, and Xinbo Gao. Multi-view matrix decomposition: a new scheme for exploring discriminative information. In *IJCAI*, 2015. 1

[6] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, pages 524–531, 2005. 5

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 4

[8] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, pages 1735–1742, 2006. 2

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2, 3

[10] Menglei Hu and Songcan Chen. Doubly aligned incomplete multi-view clustering. In *IJCAI*, page 2262–2268, 2018. 1, 2, 5, 6

[11] Peng Hu, Dezhong Peng, Yongsheng Sang, and Yong Xiang. Multi-view linear discriminant analysis network. *IEEE Transactions on Image Processing*, 28(11):5352–5365, 2019. 1

[12] Peng Hu, Xi Peng, Hongyuan Zhu, Jie Lin, Liangli Zhen, and Dezhong Peng. Joint versus independent multiview hashing for cross-view retrieval. *IEEE Transactions on Cybernetics*, 2020. 1

[13] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *CVPR*, pages 8849–8858, 2020. 4

[14] Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, and Xi Peng. Partially view-aligned clustering. In *NeurIPS*, 2020. 1

[15] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, pages 9865–9874, 2019. 4

[16] Yangbangyan Jiang, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Dm2c: Deep mixed-modal clustering. In *NeurIPS*, pages 5888–5892, 2019. 1, 2

[17] Zhao Kang, Xinjia Zhao, Chong Peng, Hongyuan Zhu, Joey Tianyi Zhou, Xi Peng, Wenyu Chen, and Zenglin Xu. Partition level multiview subspace clustering. *Neural Networks*, 122:279–288, 2020. 1

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 5

[19] Bo Li, Risheng Liu, Junjie Cao, Jie Zhang, Yu-Kun Lai, and Xiuping Liu. Online low-rank representation learning for joint multi-subspace recovery and clustering. *IEEE Transactions on Image Processing*, 27(1):335–348, 2017. 2

[20] Ruihuang Li, Changqing Zhang, Huazhu Fu, Xi Peng, Tianyi Zhou, and Qinghua Hu. Reciprocal multi-layer subspace learning for multi-view clustering. In *ICCV*, pages 8172–8180, 2019. 1

[21] Sheng Li, Ming Shao, and Yun Fu. Cross-view projective dictionary learning for person re-identification. In *IJCAI*, pages 2155–2161, 2015. 1

[22] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *AAAI*, pages 1968–1974, 2014. 1, 2, 5, 6

[23] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *AAAI*, 2021. 2

[24] Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *AAAI*, pages 2750–2756, 2015. 5

[25] Weiwei Liu, Xiaobo Shen, and Ivor Tsang. Sparse embedded k-means clustering. In *NeurIPS*, pages 3319–3327, 2017. 5

[26] 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*, 2020. 2, 5, 6

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 8

[28] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pages 6707–6717, 2020. 2

[29] Feiping Nie, Jing Li, Xuelong Li, et al. Self-weighted multiview clustering with multiple graphs. In *IJCAI*, pages 2564–2570, 2017. 1

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 2, 3

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8026–8037, 2019. 5

[32] Xi Peng, Jiashi Feng, Shijie Xiao, Wei-Yun Yau, Joey Tianyi Zhou, and Songfan Yang. Structured autoencoders for subspace clustering. *IEEE Transactions on Image Processing*, 27(10):5076–5086, 2018. 5

[33] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *ICML*, pages 5092–5101, 2019. 1

[34] Xi Peng, Hongyuan Zhu, Jiashi Feng, Chunhua Shen, Haixian Zhang, and Joey Tianyi Zhou. Deep clustering with sample-assignment invariance prior. *IEEE Transactions on*

Neural Networks and Learning Systems, 31(11):4857–4868, 2020. 4

[35] Weixiang Shao, Lifang He, and S Yu Philip. Multiple incomplete views clustering via weighted nonnegative matrix factorization with $l_{2,1}$ regularization. In *ECML PKDD*, pages 318–334, 2015. 1, 2

[36] Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. In *COLT*, pages 403–414, 2008. 2

[37] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv:1906.05849*, 2019. 2

[38] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv:2006.05576*, 2020. 2

[39] Hao Wang, Linlin Zong, Bing Liu, Yan Yang, and Wei Zhou. Spectral perturbation meets incomplete multi-view data. In *IJCAI*, pages 3677–3683, 2019. 2, 5, 6

[40] Qi Wang, Mulin Chen, Feiping Nie, and Xuelong Li. Detecting coherent groups in crowd scenes by multiview clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):46–58, 2018. 1

[41] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Partial multi-view clustering via consistent gan. In *ICDM*, pages 1290–1295, 2018. 1, 2

[42] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *ICML*, pages 1083–1092, 2015. 5, 6

[43] 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 *AAAI*, pages 5393–5400, 2019. 5, 6

[44] Jie Wen, Zheng Zhang, Yong Xu, Bob Zhang, Lunke Fei, and Guo-Sen Xie. Cdimc-net: Cognitive deep incomplete multiview clustering network. In *IJCAI*, pages 3230–3236, 2020. 2

[45] Cai Xu, Ziyu Guan, Wei Zhao, Hongchang Wu, Yunfei Niu, and Beilei Ling. Adversarial incomplete multi-view clustering. In *IJCAI*, pages 3933–3939, 2019. 1

[46] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view learning with incomplete views. *IEEE Transactions on Image Processing*, 24(12):5812–5825, 2015. 1

[47] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view self-paced learning for clustering. In *IJCAI*, pages 3974–3980, 2015. 1

[48] Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu, and Xi Peng. Partially view-aligned representation learning with noise-robust contrastive loss. In *CVPR*, 2021. 5

[49] Yingzhen Yang, Jiashi Feng, Nebojsa Jojic, Jianchao Yang, and Thomas S Huang. ℓ^0-sparse subspace clustering. In *ECCV*, pages 731–747, 2016. 5

[50] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 5

[51] Changqing Zhang, Yeqing Liu, and Huazhu Fu. Ae2-nets: Autoencoder in autoencoder networks. In *CVPR*, pages 2577–2585, 2019. 5, 6

[52] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1774–1782, 2019. 5, 6

[53] Handong Zhao, Hongfu Liu, and Yun Fu. Incomplete multimodal visual data grouping. In *IJCAI*, pages 2392–2398, 2016. 1, 2, 5, 6

[54] Runwu Zhou and Yi-Dong Shen. End-to-end adversarial-attention network for multi-modal clustering. In *CVPR*, pages 14619–14628, 2020. 1

[55] Junyan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 2, 4

Supplementary Material for
COMPLETER: Incomplete Multi-view Clustering via Contrastive Prediction

Yijie Lin[1], Yuanbiao Gou[1], Zitao Liu[2], Boyun Li[1], Jiancheng Lv[1], Xi Peng[1*]

[1] College of Computer Science, Sichuan University, China

[2] TAL Education Group, China

{linyijie.gm,gouyuanbiao,zitao.jerry.liu,liboyun.gm,pengx.gm}@gmail.com; lvjiancheng@scu.edu.cn

1. Introduction

In this supplementary material, we provide additional information including mathematical notations, mathematical derivation of our loss, network architectures, and implementation details. To investigate the effectiveness of our method, we also conduct some additional experimental analysis.

2. Notations and Definitions.

In this section, we summarize the mathematical notations used throughout the manuscript in Table 1 for a clear reference. The bi-view dataset $\bar{\mathbf{X}}$ consists of three parts, *i.e.*, $\bar{\mathbf{X}}^{1,2}$, $\bar{\mathbf{X}}^{1}$, and $\bar{\mathbf{X}}^{2}$, where $\bar{\mathbf{X}}^{1,2}$, $\bar{\mathbf{X}}^{1}$, and $\bar{\mathbf{X}}^{2}$ denote the examples presented in all views, the first view only, and the second view only, respectively. n and m denote the number of the data points of the whole dataset $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}^{1,2}$, respectively. More specifically, Fig. 1 visually illustrates our setting and some notations by taking a dataset as a showcase.

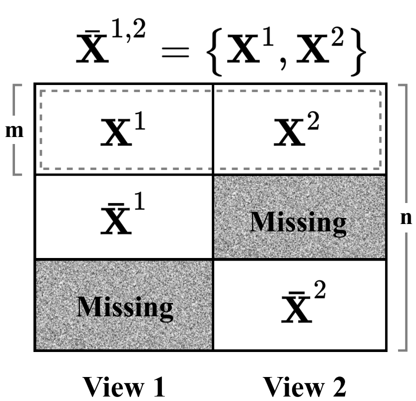


Figure 1. Illustrations of the incomplete bi-view dataset $\bar{\mathbf{X}}$.

Notation	Definition
$\bar{\mathbf{X}}$	Incomplete bi-view dataset where $\bar{\mathbf{X}} = \{\bar{\mathbf{X}}^{1,2}, \bar{\mathbf{X}}^{1}, \bar{\mathbf{X}}^{2}\}$.
$\bar{\mathbf{X}}^{1,2}$	Set of examples presented in both views.
$\bar{\mathbf{X}}^{v}$	Set of examples only presented in view v.
$\mathbf{X}^{v}$	The v-th view of complete samples $\bar{\mathbf{X}}^{1,2}$, *i.e.*, $\bar{\mathbf{X}}^{1,2} = \{\mathbf{X}^1, \mathbf{X}^2\}$.
$\mathbf{Z}^{v}$	The representations of $\mathbf{X}^{v}$.
n	Number of examples presented in $\bar{\mathbf{X}}$.
m	Number of examples presented in $\bar{\mathbf{X}}^{1,2}$.
α	Trade-off parameter of information entropy.
λ_1	Trade-off parameter of dual prediction loss.
λ_2	Trade-off parameter of reconstruction loss
$f^{(v)}$	Encoder of v-th view.
$g^{(v)}$	Decoder of v-th view.
$G^{(i)}$	Predictor which recovers missing representation $\mathbf{Z}^{j}$ from complete one $\mathbf{Z}^{i}$.

Table 1. Mathematical notations in the manuscript.

3. Theoretical Derivation on Our Loss

In this section, we elaborate on the derivations of our loss that are omitted in the manuscript due to the space limitation.

3.1. Cross-view Contrastive Loss

We have imported the information entropy to the standard definitions of mutual information in the manuscript. Mathematically, the contrastive loss is defined as,

$$\mathcal{L}_{cl} = -\sum_{t}^{m} \left(I\left(\mathbf{Z}_t^1, \mathbf{Z}_t^2\right) + \alpha \left(H\left(\mathbf{Z}_t^1\right) + H\left(\mathbf{Z}_t^2\right) \right) \right). \quad (1)$$

As illustrated in the manuscript, the representations of t-th sample $\mathbf{Z}_t^1 \in \mathbb{R}^D$ and $\mathbf{Z}_t^2 \in \mathbb{R}^D$ can be interpreted as the distribution of discrete random variables z and z' over D classes, respectively. In other words, the probability distributions $\mathcal{P}(z = d)$ is the d-th element of $\mathbf{Z}_t^1$ where

*Corresponding author

$1 \le d \le D$. Hence, considering the case of a data set or batch, the joint probability distribution $\mathcal{P}(z, z')$ of variable z and z' could be defined by $\mathbf{P} \in \mathcal{R}^{D \times D}$, *i.e.*,

$$\mathbf{P} = \frac{1}{m} \sum_{t=1}^{m} \mathbf{Z}_t^1 \left(\mathbf{Z}_t^2\right)^\top . \tag{2}$$

Thus, for discrete distributions, the mutual information and information entropy are given as below:

$$\begin{aligned} I(\mathbf{Z}^1, \mathbf{Z}^2) &= \sum_{z=1}^{D} \sum_{z'=1}^{D} \mathcal{P}(z, z') \log \left(\frac{\mathcal{P}(z, z')}{\mathcal{P}(z)\mathcal{P}(z')} \right) \\ &= \sum_{d=1}^{D} \sum_{d'=1}^{D} \mathbf{P}_{dd'} \ln \frac{\mathbf{P}_{dd'}}{\mathbf{P}_d \cdot \mathbf{P}_{d'}}, \end{aligned} \tag{3}$$

$$\mathrm{H}(\mathbf{Z}^1) = -\sum_{z=1}^{D} \mathcal{P}(z) \log \mathcal{P}(z) = -\sum_{d=1}^{D} \mathbf{P}_d \ln \mathbf{P}_d, \tag{4}$$

$$\mathrm{H}(\mathbf{Z}^2) = -\sum_{z'=1}^{D} \mathcal{P}(z') \log \mathcal{P}(z') = -\sum_{d'=1}^{D} \mathbf{P}_{d'} \ln \mathbf{P}_{d'}, \tag{5}$$

where $\mathbf{P}_d$ and $\mathbf{P}'_d$ denote the marginal probability distributions $\mathcal{P}(z = d)$ and $\mathcal{P}(z' = d')$ which could be obtained by summing over the d-th row and d'-th column of $\mathbf{P}$, respectively. By substituting Eq. (3), Eq. (4), and Eq. (5) into Eq. (1), we could obtain the final form of our cross-view contrastive loss as below:

$$\begin{aligned} \mathcal{L}_{cl} &= -\left(I\left(\mathbf{Z}^1, \mathbf{Z}^2\right) + \alpha \left(H\left(\mathbf{Z}^1\right) + H\left(\mathbf{Z}^2\right)\right)\right) \\ &= -\left(\sum_{d=1}^{D} \sum_{d'=1}^{D} \mathbf{P}_{dd'} \ln \frac{\mathbf{P}_{dd'}}{\mathbf{P}_d \cdot \mathbf{P}_{d'}} - \right. \\ &\qquad \left. \alpha \left(\sum_{d=1}^{D} \mathbf{P}_d \ln \mathbf{P}_d + \sum_{d'=1}^{D} \mathbf{P}_{d'} \ln \mathbf{P}_{d'} \right) \right) \\ &= -\left(\sum_{d=1}^{D} \sum_{d'=1}^{D} \mathbf{P}_{dd'} \ln \frac{\mathbf{P}_{dd'}}{\mathbf{P}_d \cdot \mathbf{P}_{d'}} + \right. \\ &\qquad \left. \alpha \left(\sum_{d=1}^{D} \sum_{d'=1}^{D} \mathbf{P}_{dd'} \ln \frac{1}{\mathbf{P}_d} + \sum_{d'=1}^{D} \sum_{d=1}^{D} \mathbf{P}_{dd'} \ln \frac{1}{\mathbf{P}_{d'}} \right) \right) \\ &= -\sum_{d=1}^{D} \sum_{d'=1}^{D} \mathbf{P}_{dd'} \left(\ln \frac{\mathbf{P}_{dd'}}{\mathbf{P}_d \cdot \mathbf{P}_{d'}} + \alpha \left(\ln \frac{1}{\mathbf{P}_d} + \ln \frac{1}{\mathbf{P}_{d'}} \right) \right) \\ &= -\sum_{d=1}^{D} \sum_{d'=1}^{D} \mathbf{P}_{dd'} \ln \frac{\mathbf{P}_{dd'}}{\mathbf{P}_d^{\alpha+1} \cdot \mathbf{P}_{d'}^{\alpha+1}}. \end{aligned} \tag{6}$$

3.2. Cross-view Dual Prediction Loss

To infer the missing views $\mathbf{Z}^i$ from $\mathbf{Z}^j$, we propose to minimize the conditional entropy $H(\mathbf{Z}^i|\mathbf{Z}^j) = -\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}[\log \mathcal{P}(\mathbf{Z}^i|\mathbf{Z}^j)]$. As it is intractable to solve such a problem, we introduce a variational distribution $\mathcal{Q}(\mathbf{Z}^i|\mathbf{Z}^j)$ and further maximize the lower bound of $\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}[\log \mathcal{P}(\mathbf{Z}^i|\mathbf{Z}^j)]$, *i.e.*, $\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}\left[\log \mathcal{Q}\left(\mathbf{Z}^i \mid \mathbf{Z}^j\right)\right]$. To be specific, we assume $\mathcal{Q}$ as a Gaussian distribution $\mathcal{N}\left(\mathbf{Z}^i \mid G^{(j)}\left(\mathbf{Z}^j\right), \sigma \mathbf{I}\right)$, where $G^{(j)}(\cdot)$ is the predictor which maps $\mathbf{Z}^j$ to $\mathbf{Z}^i$ and $\sigma\mathbf{I}$ is a variance matrix. As a result, we have

$$\begin{aligned} &\max \mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}\left[\log \mathcal{Q}\left(\mathbf{Z}^i \mid \mathbf{Z}^j\right)\right] = \\ &\qquad \mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}\left[\log\left(\frac{1}{\sqrt{\sigma I}\sqrt{2\pi}} e^{-\frac{\left(\mathbf{z}^i - G^{(j)}\left(\mathbf{z}^j\right)\right)^2}{2(\sigma \mathbf{I})}}\right)\right], \end{aligned} \tag{7}$$

which could be equivalent to the following optimization problem:

$$\max \mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}}\left[-\frac{\left(\mathbf{Z}^i - G^{(j)}\left(\mathbf{Z}^j\right)\right)^2}{2\sigma\mathbf{I}} + \log \frac{1}{\sqrt{2\pi\sigma I}}\right]. \tag{8}$$

By ignoring the constant $\log \frac{1}{\sqrt{2\pi\sigma I}}$ and scaling factor $2\sigma\mathbf{I}$, we could obtain the prediction loss as below,

$$\max -\mathbb{E}_{\mathcal{P}_{\mathbf{Z}^i,\mathbf{Z}^j}} \left\| \left(\mathbf{Z}^i - G^{(j)}\left(\mathbf{Z}^j\right)\right) \right\|_2^2 . \tag{9}$$

Hence, for a given bi-view dataset, the dual prediction loss is accordingly defined as

$$\mathcal{L}_{pre} = \left\| G^{(1)}(\mathbf{Z}^1) - \mathbf{Z}^2 \right\|_2^2 + \left\| G^{(2)}(\mathbf{Z}^2) - \mathbf{Z}^1 \right\|_2^2 . \tag{10}$$

4. Experiment Details

In this section, we elaborate on the implementation details of our method and the experimental settings.

4.1. Network Architectures of COMPLETER

The proposed method contains two training modules, *i.e.*, view-specific auto-encoders and cross-view prediction networks. Table 2 and 3 have presented the details of the network architectures in these two training modules, respectively. For these two modules, we simply adopt a dense (*i.e.*, fully-connected) network where each layer is followed by a batch normalization layer and a ReLU layer. The softmax activation function is used at the last layer of the encoders and prediction modules. The structures of autoencoders $f^{(\cdot)}$, $g^{(\cdot)}$, and predictors $G^{(\cdot)}$ for different views are the same. Specifically, the size of the output of the encoder and predictor should be the same and set to 64 or 128 according to the dataset.

4.2. Implementation Details for Clustering

To perform clustering, we adopt k-means to compute the cluster assignments on the common representation which is

Table 2. The architecture of the autoencoders in COMPLETER.

Dataset	Encoder	Decoder
Caltech101-20	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (Softmax, size = **128**)	Dense (BatchNorm1d, ReLU, size = input)
LandUse-21	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (Softmax, size = **64**)	Dense (BatchNorm1d, ReLU, size = input)
Scene-15	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (Softmax, size = **128**)	Dense (BatchNorm1d, ReLU, size = input)
Noisy MNIST	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (BatchNorm1d, ReLU, size = 1024)	Dense (BatchNorm1d, ReLU, size = 1024)
	Dense (Softmax, size = **64**)	Dense (BatchNorm1d, ReLU, size = input)

Table 3. The architecture of dual prediction in COMPLETER.

Structure
Dense (BatchNorm1d, ReLU, size = 128)
Dense (BatchNorm1d, ReLU, size = 256)
Dense (BatchNorm1d, ReLU, size = 128)
Dense (BatchNorm1d, ReLU, size = 256)
Dense (BatchNorm1d, ReLU, size = 128)
Dense (Softmax, size = input)

Table 4. Influence of dimensionality.

Dataset	Dimension	ACC	NMI	ARI
Caltech101-20	32	43.48	60.31	41.50
	64	51.99	62.88	47.91
	128	68.44	**67.39**	**75.44**
	256	**69.56**	65.63	74.54
Scene-15	32	37.30	40.79	21.41
	64	37.60	41.01	20.55
	128	**39.50**	**42.35**	**23.51**
	256	36.37	41.87	22.10

obtained by simply concatenating all view-specific representations together. Specifically, we use the k-means contained in the Scikit-Learn package [3] with the default configuration. For a fair comparison, we run all the used methods five times with different initializations and data partitions to obtain the common representations. For each run, we further conduct k-means 10 times on the representation to obtain the clustering results. In all experiments, we adopt three evaluation metrics implemented by Scikit-Learn to evaluate the clustering performance, namely, ACC, NMI, and ARI.

5. Additional Experiments

This section presents two experimental studies including: i) the influence of dimensionality of latent representations and ii) clustering performance on the full datasets.

5.1. Influence of Dimensionality

In the proposed method, we treat each element of the representation as an over-cluster class probability like [1, 2, 4]. To evaluate the effectiveness of such over-clustering strategy, we change the dimensionality of the representation in the range of $\{32, 64, 128, 256\}$. The missing rate η is fixed to 0.5 and the results are shown in Table 4. The results demonstrate that a too large or too small dimensionality will cause performance degradation. The former is of insufficient representability and the latter may have some redundant information.

5.2. Experiment on the Full Datasets

In the main body of the manuscript, we only report the results on the 20k subsets of the Noisy MNIST dataset because most of the baselines are inefficient to handle large scale datasets. In this evaluation, we carry out clustering experiments on the whole Noisy MNIST dataset and report the results compared with scalable methods including DCCA [5], DCCAE [5], BMVC [7], and AE^{2}Nets [6]. Similarly, we also test these methods in both *Incomplete* ($\eta = 0.5$) and *Complete* ($\eta = 0$) settings. As shown in Table 5, COMPLETER still outperforms all baselines.

Table 5. Performance comparisons on full Noisy MNIST.

Missing Type	Method	ACC	NMI	ARI
Incomplete	DCCA	45.32	48.73	25.70
	DCCAE	49.44	48.49	25.31
	AE^{2}Nets	37.76	35.53	20.57
	BMVC	46.42	36.23	22.34
	COMPLETER	**94.28**	**87.39**	**88.12**
Complete	DCCA	89.29	91.35	87.04
	DCCAE	89.03	91.40	87.77
	AE^{2}Nets	50.70	53.26	40.49
	BMVC	91.57	83.55	83.83
	COMPLETER	**97.17**	**94.19**	**93.58**

References

[1] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *CVPR*, pages 8849–8858, 2020. 3

[2] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, pages 9865–9874, 2019. 3

[3] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 3

[4] Xi Peng, Hongyuan Zhu, Jiashi Feng, Chunhua Shen, Haixian Zhang, and Joey Tianyi Zhou. Deep clustering with sample-assignment invariance prior. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11):4857–4868, 2020. 3

[5] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *ICML*, pages 1083–1092, 2015. 3

[6] Changqing Zhang, Yeqing Liu, and Huazhu Fu. Ae2-nets: Autoencoder in autoencoder networks. In *CVPR*, pages 2577–2585, 2019. 3

[7] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1774–1782, 2019. 3