Deep Distribution-preserving Incomplete Clustering with Optimal Transport

Anonymous ICCV submission

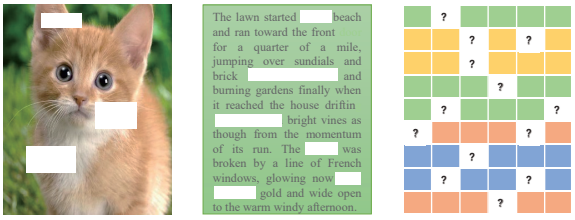
Paper ID 4151

Abstract

Clustering is a fundamental task in the computer vision and machine learning community. Although various methods have been proposed, the performance of existing approaches drops dramatically when handling incomplete high-dimensional data (which is common in real world applications). To solve the problem, we propose a novel deep incomplete clustering method, named Deep Distribution-preserving Incomplete Clustering with Optimal Transport (DDIC-OT). In the proposed algorithm, an auto-encoder is utilized to reconstruct the sample and impute the missing values. Different from the common auto-encoders, to better impute the missing values and learn a more discriminative latent representation for clustering, two mechanisms are proposed. First, distribution distance measured by the optimal transport is introduced for reconstruction evaluation instead of the pixel-wise loss function. Second, the clustering loss of the latent feature is designed to regularize the embedding with more discrimination capability. As a consequence, the network becomes more robust against absent features and noise within samples. Also, the unified framework which combines clustering and sample imputation enables the two procedures to negotiate to better serve for each other. Extensive experiments demonstrate that the proposed network achieves superior and stable clustering performance improvement against existing state-of-the-art incomplete clustering methods over different missing ratios.

1. Introduction

Clustering is one of the fundamental and important unsupervised learning tasks in data science, image analysis and machine learning community [15, 14, 21, 27, 9, 33, 16, 29, 32]. A wide variety of data clustering methods have been proposed to organise similar items into same groups and achieve promising performance, e.g., k-means clustering, Gaussian Mixture Model (GMM), spectral clustering and deep clustering recently. However, existing clustering approaches all hold one premise that the data themselves are complete while data with missing features are quite common in reality. Data incompleteness occurs due to many factors, e.g. sensor failure, unfinished collection and data storage corruption. For example, face images are covered with masks leading to missing features during COVID-19 [5]. When facing with various types of missing features, incomplete data clustering has draw increasing attention in recent years [24, 13, 25, 30, 15, 14](see Figure. 1).



(a) Masked Images (b) Obstructed Text (c)Matrix Form

Figure 1: (a) Example image is presented with missing features due to obstruction or shadow [11];(b) Text data may be occupied with imperfect storage.(c) All the numerical incompleteness phenomenons can be represented with the matrix form. The data are split with two counterparts: observed and missing ones.

Existing incomplete clustering can be roughly categorized into two mechanisms, **heuristic-based** and **learning-based** respectively. Both of them firstly impute the missing features and then the full data matrix can be applied with traditional clustering algorithms. The heuristic imputation methods often rely on statistic property, e.g., zero-filling (ZF) and mean-filling (MF) after normalizing. Median values are also popular for imputation in genetic study. Particularly, the KNN-filling method considers the local reliable partners which fills the missing entries with the mean value of k-closest neighbors. When facing with complex high-dimensional data, heuristic-based methods perform poorly since the simple imputations cannot obtain enough information to precisely recover data.

Recently, learning-based imputation methods receive enormous attention and become to be the mainstream. Existing work can be categorized into shallow and deep learn-

ing framework. The shallow representatives normally assume that the data are low-rank and therefore apply iterative methods to recover missing values [20, 8, 26, 17, 3]. Moreover, the Expectation-Maximum (EM) algorithm iteratively estimates the maximum likelihood and then inferences the missing variables until convergent. With the improvements of deep learning architectures, various deep networks have been proposed to handle incompleteness. A desirable attribute for deep approaches is that they should accurately inference the joint and marginal distributions of the data. Therefore, variants of generation-style networks are introduced including Generative Adversarial Networks (GAN) and Variational Auto-Encoder (VAE). In [31], a generator utilizes the observed features to generate 'complete' data and the discriminator attempts to determine which components are actually observed or imputed. With the adversary training strategy, the generated missing features could approximate the real data distribution. Followed this line, enormous GAN and VAE-based approaches are put forward to minimizing the distances between real values and imputed matrices.

Although these aforementioned methods offer solutions for incomplete data clustering, several drawbacks in existing mechanism cannot be neglected: i) Existing incomplete clustering methods follow a two-step manner, where the imputation stage and the clustering stage are separated from each other. In other words, the imputed features are not designed for clustering task, which may heavily degrade the clustering performance in return. ii)When facing with high-dimensional data (e.g., images, text), both of the shallow and deep methods perform poorly due to the insufficient observed information with inaccurate imputation. These results in sharp degradation in clustering task performance.

In this paper, we propose a novel deep incomplete clustering method, which we refer as Deep Distribution-preserving Incomplete Clustering with Optimal Transport (DDIC-OT), that generalizes the well-known Deep Embedding Clustering network (DEC) handle missing features. Different from existing pixel-by-pixel reconstruction in traditional autoencoder, we propose to minimize the Wasserstain distance between observed data and the reconstructed data with optimal transport. Moreover an addition clustering layer is added into the embedded representation level with KL-divergence for measuring clustering loss. By optimizing the novel network, the distribution of original data can be well-preserved and in return the missing features can be more accurately imputed by guidance of latent clustering structures. Thus, the proposed DDIC-OT simultaneously utilizes the imputation and the embedded clustering procedures so that they can be jointly negotiated with each other and reach consensus best serving for clustering task. Finally, the proposed DDIC-OT is showcased in extensive experiments on a wide variety of benchmarks with

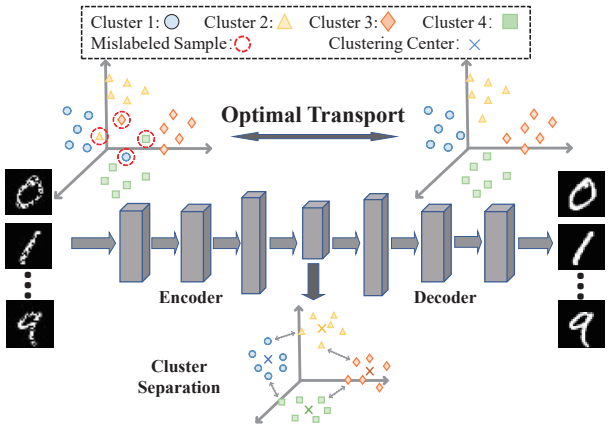


Figure 2: The framework of the proposed DDIC-OT. Instead of pixel-to-pixel reconstruction, we impute features by minimizing the distribution distance with optimal transport. Moreover, the latent embedding representations are regularized with clustering loss to ensure intra-cluster discrimination. The joint loss functions seamlessly negotiate incomplete imputation and clustering tasks into a unified framework contributes to clustering improvements.

different missing ratios, to evaluate its effectiveness. As demonstrated, the proposed network enjoys superior clustering performance in comparison with existing state-of-the-art imputation methods by large margins.

The contributions of our DDIC-OT are summarized as follows,

1. We mathematically analyze the failures of existing incomplete clustering methods in theory when facing with high-dimensional data. To avoid insufficient training brought by the spareness of full-observed data, a novel end-to-end deep clustering network is proposed to minimize the wasserstain distances between original and reconstructed distribution.

2. We regularize the latent distribution with more discriminate separation to further enhance task performance. By the guidance of the unified loss function, the network decodes the informative latent representations contributing to better recovery and clustering. To the best of our knowledge, this could be the first work of end-to-end deep incomplete clustering network.

3. Comprehensive experiments are conducted on six high-dimensional benchmarks datasets with various incomplete ratios. As the experimental results show, the proposed network significantly outperforms state-of-the-art incomplete clustering methods by large margins.

2. Notion and Related Work

2.1. Notion

Throughout this paper, we use boldface uppercase letters and lowercase letters to denote matrices and vectors respectively. The (i, j)-th elements of a matrix $\mathbf{U}$ is referred as $\mathbf{U}_{ij}$.

Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the data matrix with n samples d dimension. Then we define a missing index matrix $\mathbf{M} \in \{0, 1\}^{n\times d}$ as follows,

$$\mathbf{M}_{ij} = \begin{cases} 1 & \text{if the entry } (i, j) \text{ of } \mathbf{X} \text{ can be observed,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

With the defined mask matrix $\mathbf{M}$, the incomplete data matrix we observed can be defined as

$$\mathbf{X}^{(obs)} = \mathbf{X} \circ \mathbf{M} + NaN \circ (\mathbf{1}_{n\times d} - \mathbf{M}), \tag{2}$$

where the $\circ$ denotes the Hadamard (elementwise) product and NaN means **N**ot **a** **N**umber throughout the paper.

2.2. Related Work

Statistical imputation. Basic statistical methods try to utilize information from the missing data by the means of numerical property. Most of them use statistical attributes to estimate the missing feature values, rather than directly discard incomplete feature information. Incomplete entries are filled with constants to obtain complete data so that they can be directly applied to machine learning tasks, e.g., zero, mean and median. Additionally, KNN imputation method has been considered as an alternative estimating the missing features with the mean of k nearest reliable neighbors [1].

The Bayesian framework is different from the previous method in that it considers the joint and conditional distribution for dealing with incomplete features. These frames are generally expressed in terms of a maximum-likelihood method, which estimate missing values with the most probable numbers. The most popular method of Bayesian framework is the Expectation Maximization (EM) algorithm [2, 4].

Deep incomplete clustering. Although deep clustering mechanism has received much attention in recent years, none of existing methods has considered to cluster with incomplete features in an end-to-end manner yet. The typical methods follow the two-step strategy:imputation and clustering separately. They propose to fill missing values through neural networks and then apply clustering algorithms on the estimated dataset. GAIN [31] firstly proposes to impute incomplete features with GAN. Different from the traditional GAN networks, the goal of the discriminator in GAIN is to accurately distinguish whether the data are imputed or real, so as to force the samples generated by the generator to be close to the real data distribution. Unfortunately, the same problems as common GANs, these models are generally difficult to train since the optimization processes are hardly stable. Apart from GAN-like networks, VAEAC [7] proposes a neural probabilistic model based on variational autoencoder, which can estimate the observed features using stochastic gradient variational inference [10]. However these VAE-based method may lead to poor results when the posterior approximation of variational inference is far from the actual posterior approximation. In addition, based on fitting the conditional distribution of the missing data, a Markov chain Monte Carlo (MCMC) scheme has been developed in [23].

Optimal transport and sinkhorn divergence. Let $\alpha = \sum_{i=1}^{n} a_i \delta_{\mathbf{X}_i}$, $\beta = \sum_{i=1}^{n} b_i \delta_{\mathbf{Y}_i}$ be two discrete distributions formed by empirical given data samples $\mathbf{X}, \mathbf{Y}$, and their supports $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{Y} \in \mathbb{R}^{n'\times d}$ and frequency vectors $\mathbf{a}, \mathbf{b}$. It can be easily obtained that $\mathbf{a}^\top \mathbf{1} = 1, \mathbf{a} \geq 0, \mathbf{b}^\top \mathbf{1} = 1, \mathbf{b} \geq 0$.The q-th Wasserstein distance corresponds to these two distributions α and β is denoted as follows,

$$W_q(\alpha, \beta) \overset{\text{def}}{=} \min_{\mathbf{P}\in U(\mathbf{a},\mathbf{b})} \langle \mathbf{F}, \mathbf{C} \rangle, \tag{3}$$

where $U(\mathbf{a}, \mathbf{b}) \overset{\text{def}}{=} \left\{ \mathbf{F} \in \mathbb{R}_{+}^{n\times n'} : \mathbf{F}\mathbf{1} = \mathbf{a}, \mathbf{F}^\top \mathbf{1}_n = \mathbf{b} \right\}$ and $\mathbf{C} = (\|x_i - y_j\|^q)_{ij} \in \mathbb{R}^{n\times n'}$ denotes as the cost matrix of pairwise squared distances between the support sets. In our paper, we set $q = 2$. The Wasserstein distance denoted in Eq. (3) is often jointly introduced with an entropy regularization,

$$W_q^{\epsilon}(\alpha, \beta) \overset{\text{def}}{=} \min_{\mathbf{F}\in U(\mathbf{a},\mathbf{b})} \langle \mathbf{F}, \mathbf{C} \rangle - \epsilon h(\mathbf{F}), \tag{4}$$

where $h(\mathbf{F}) \overset{\text{def}}{=} -\sum_{ij} f_{ij} \log f_{ij}$ denotes the entropy regularization. Eq. (4) can be efficiently optimized using Sinkhorn algorithm [22]. Based on Eq. (4), a symmetric divergence can be represented as,

$$S_\epsilon(\alpha, \beta) \overset{\text{def}}{=} \mathrm{OT}_\epsilon(\alpha, \beta) - \frac{1}{2} \left(\mathrm{OT}_\epsilon(\alpha, \alpha) + \mathrm{OT}_\epsilon(\beta, \beta) \right). \tag{5}$$

The Sinkhorn divergence in Eq. (5) offers an tractable alternative for Wasserstein distance calculations, and easily be accelerated by GPU. In our paper, we use the sinkhorn divergence to measure the OT distance of two distributions.

3. DDIC-OT

3.1. Motivation

Problem analysis Although the aforementioned methods have been proposed to solve incomplete data clustering to some extent, most of them are evaluated with very small-dimensional data and make them unpractical in real scenarios. When facing with high-dimensional data (e.g., images,

text), both of the existing shadow and deep methods perform poorly due to the insufficient observed information with inaccurate imputation. We theoretically analyze this phenomenon with the following Theorem 1.

Theorem 1 *Suppose the data are i.i.d (independently and identically distributed), a fully-observed high-dimensional data sample exists with low probability when facing incompleteness.*

Proof 1 *Suppose the missing ratio is* $p(0 \leq p \leq 1)$. *Given a matrix* $\mathbf{X} \in \mathbb{R}^{n \times d}$. *For each sample* x_i, *we can obtain the following equation.*

$$P\left(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{id}\right) = P\left(\mathbf{X}_{i1}\right) \cdots P\left(\mathbf{X}_{id}\right) = (1-p)^d, \tag{6}$$

where $P\left(\mathbf{X}_{i1}\right)$ *denotes the probability* $\mathbf{X}_{i1}$ *can be observed.*

Taking $p = 0.1, d = 300$ *as example, with* 10^{-14} *probability the sample* x_i *can be fully-observed. With the increasing dimension* d, *the probability becomes smaller and approximates 0. This completes the proof.*

The Theorem 1 illustrates that very few samples are fully complete when the dimensions are relatively high. Therefore, the traditional statistical and deep generative methods fail to impute proper values lacking of sufficient information, e.g., knn-filling and GAN-style solutions. In [19], the Wasserstain distance is firstly applied to impute missing features where the assumption is to minimize the discrepancy of missing data distribution and the complete data distribution. In the low-dimensional incomplete setting (less than 50), the experiments results are promising and show more stable consistency in respect to the incomplete ratio change. However, as confronting with much higher dimension data types (e.g.,images,videos and text), very few fully-complete data can be obtained making the empirical estimation of target distribution (complete data distribution) difficult and inaccurate. Therefore these methods perform poorly in downstream clustering tasks (see results in Table. 1).

Different from existing assumptions, we propose to jointly solve the two processes in a unified framework: reconstruction and clustering. The natural way of reconstruction is to apply autoencoder models. The observed values can be regraded as 'supervised' signals for the reconstruction. However, with few informative information, it is not reasonable to only reconstruct the missing counterparts since the reconstruction may destroy the geometry distribution features for the data and the clustering performance is heavily affected. In this paper, we decide to recover the latent distribution instead of pixel-level approximation. Specially, we adopt the latent variable models defined by an encoder-decoder manner, where we firstly encode original data x_i into the latent code z_i in the latent space $\mathcal{Z}$ and then z_i is decoded to the reconstructed image $\hat{x}_i$. This process can be expressed as,

$$p_{\hat{X}}(\hat{x}) := \int_{\mathcal{Z}} p_{p_{\hat{X}}(\hat{x}|z)}(\hat{x}|z) p_z(z) dz, \quad \forall x \in \mathcal{X} \tag{7}$$

where $p_x(z|x), p_{\hat{X}}(\hat{x}|z)$ are parameterized with the encoder f_e and decoder f_d network. Then the distribution-preserving loss can be measured with Eq. (5) respecting to p_X and $p_{\hat{X}}$,

$$L_s(\mathbf{X}, \hat{\mathbf{X}}) = S_\epsilon(\mathbf{X}, f_d(f_e(\mathbf{X}))). \tag{8}$$

3.2. Overall Network Architecture

In this section, we leverage the one-stage deep incomplete clustering introduced in the previous section as a basis to demonstrate the process of the proposed learning algorithm, the overall flowchart is illustrated in Figure 2. The proposed clustering model consists of three parts, an encoder, a decoder, and a soft clustering layer, specifically, the method relies on a linear combination based on two objective functions, representing the optimal transport distance and clustering loss respectively. The joint optimization process can be described as follows:

$$L = L_s + \gamma L_c, \tag{9}$$

where L_s is the sinkhorn divergence shown in Eq. (5) and L_c is the clustering loss. γ is a hyper-parameter, which is used to balance the two costs. Consider a dataset $\mathbf{X}$ with n samples, and each $x_i \in \mathbb{R}^d$ where d is the dimension. The number of clusters k is known, for each input data x_i we denote the nonlinear mapping $f_e : x_i \to z_i$ and $f_d : z_i \to \hat{x}_i$ where z_i is the low dimensional feature space, $\hat{x}_i$ is the complete data learned through the network.

The clustering loss is defined as KL divergence between distributions P and Q proposed in[28], where P is the soft assignment of the distribution z:

$$p_{ij} = \frac{\left(1 + \|z_i - \mu_j\|^2\right)^{-1}}{\sum_j \left(1 + \|z_i - \mu_j\|^2\right)^{-1}}, \tag{10}$$

and then the cluster assignment can be obtained $s_i = \arg\max_j p_{ij}$. Then Q is the target distribution derived from P,

$$q_{ij} = \frac{p_{ij}^2 / \sum_i p_{ij}}{\sum_j \left(p_{ij}^2 / \sum_i p_{ij}\right)}. \tag{11}$$

Therefore, the clustering loss is defined as

$$L_c = KL(Q \| P) = \sum_i \sum_j q_{ij} \log \frac{q_{ij}}{p_{ij}}. \tag{12}$$

We summarize the merits of our proposed framework with the following factors: i) more naturally handle with incomplete clustering in high-dimensional space. The L_s loss accomplishes the reconstruction samples with preserving geometry characteristics. ii) more flexible that does not require the prior distribution of $\mathbf{X}$ or $\mathbf{Z}$. Instead of explicit distribution formulation, our encoder and decoder network implicitly estimate the latent distribution with more flexibility; iii) regularizing the latent distribution q with more discriminate separation to further enhance task performance. As the empirical experimental results show, the guidance of the joint loss function updates the network leading to the improvement of clustering performance.

3.3. Model Training

The training phase of the model consists of two phases: the pre-training phase of the autoencoder, where the network only contains reconstruction loss and the fine-tune phase, where both optimal transport distance and clustering loss are optimized. The auto-encoder we used in the experiments is the same across all datasets. Our encoder is a fully-connected multi-layer perceptron with dimensions d-500-500-1000/2000-10 and the decoder is a mirrored version of the encoder. The details of the network architecture are provided in the supplementary materials. The training is based on the Adam optimizer with standard learning rate η = 0.001 and the batch size is set to 256 on all datasets. After pre-training, We provide two sets of options of 150 and 100 for the parameter λ. Furthermore, for sinkhorn divergence, we set entropy regularization parameters ϵ as 0.01 in all datasets. we will stop training if the percentage of label distribution change between two consecutive updates of the target distribution is less than the threshold δ or the epoch satisfies $MaxIter$. We set δ as 0.1, $MaxIter$ as 200 in all experiments. The whole algorithm is summarized in Algorithm 1.

Algorithm 1 DDIC-OT

Input: Missing data $\mathbf{X}_m$; Cluster number k; Hyper-parameter λ; Batchsize N; Maximum iterations $MaxIter$; Stopping threshold δ; Learning rate η.
Output: Clustering Assignment S.

Initialize $\mathbf{X}_m$ with mean filling.
Initialize clustering centroids u.
for $iter = 0$ to $Maxiter$ **do**
  for $i = 0$ to $\lfloor n/N \rfloor$ **do**
    Sample a minibatch $\{x_i\}_{i=1}^{m}$ from $\mathbf{X}$.
    Compute related variables by $z_i = f_e(x_i)$, $\hat{x_i} = f_d(z_i)$.
    Compute P_i, Q_i using Eq. (10) and Eq. (11)
    Compute clustering assignment for $\{x_i\}_{i=1}^{m}$.
    Compute overall loss L by Eq. (9).
    Back-propagation and update model weights.
  end for
  Compute $Z = f_e(\mathbf{X})$.
  Compute P and Q.
  Compute clustering assignment S.
  if $sum(S_{iter+1} \neq S_{iter})/n < \delta$ **then**
    Stop training.
  end if
end for

4. Experiments

4.1. Experiments Setup

Datasets In this paper, we conduct extensive experiments on the six widely-used large-scale benchmark datasets. (1) **MNIST-full** and **Fashion-MNIST**[12]: 70000 images including the training and testing split are combined into a unified dataset. (2)**USPS**[6]: This dataset contains a total of 9298 grayscale samples with 16 × 16 pixels. (3) **COIL-20**[1]:COIL-20 consists of 1440 images of 20 objects taken by cameras from varying angles. (4)**Reuters-10K**: We used 4 root categories: corporate/industrial, government/social, markets and economics as labels and excluded all documents with multiple labels. We randomly sampled a subset of 10000 examples and computed tf-idf features on the 2000 most frequent words. We term this dataset as Reuters-10K.(5)**Letter** [2]: The Letter dataset merges a balanced set of the 26 letters with 800 images each class.

Followed by existing incomplete clustering task setting, we set seven groups of integrity ratios as $\{0.1, 0.2, 0.3, \cdots 0.6, 0.7\}$ for each dataset in our experiments. Integrity ration means the percentage of missing features in all samples.

Evaluation Metrics In our experiements, we used three standard clustering performance metrics for evaluation: (1) **Accuracy (ACC)** is computed by assigning each cluster with the dominating class label and taking the average correct classification rate as the final score, (2) **Normalised Mutual Information (NMI)** quantifies the normalised mutual dependence between the predicted labels and the ground-truth, and (c) **Purity** measures the proportion of the number of samples correctly clustered to the total number of samples. All of these metrics scale from 0 to 1 and higher values indicate better performance. Specially, we report the mean values and standard derivations of 10 independent runs to avoid the randomness brought by the different initializations of k-means.

4.2. Compared SOTA methods

(1) **Mean-Filling (MF)**: The missing features are imputed with the mean of the observed values in the correspond-

[1]https://www.cs.columbia.edu/CAVE/software/softlib/

[2]https://www.nist.gov/itl/products-and-services/emnist-dataset

Table 1: The aggregated ACC, NMI and Purity comparison (mean±std) of different algorithms on benchmark datasets. '-' means out of the GPU memory. The detailed results are omitted due to space limit and provided in supplementary materials.

Method	Shallow					Deep				
Dataset	MF	ZF	LRC	MNC	FSGR	GAIN	VAEAC	MIWAE	MDIOT	Ours
ACC(%)										
Mnist	54.66 ± 3.13	52.48 ± 3.33	51.82 ± 3.46	53.28 ± 3.08	53.22 ± 2.97	52.05 ± 3.06	53.73 ± 3.63	-	54.31 ± 3.46	$\mathbf{84.31 \pm 0.0}$
Usps	61.63 ± 3.92	60.72 ± 3.24	61.78 ± 3.63	61.80 ± 3.61	60.22 ± 3.38	60.20 ± 3.11	61.90 ± 3.51	48.75 ± 4.73	63.63 ± 4.07	$\mathbf{74.48 \pm 0.0}$
Fmnist	51.14 ± 3.87	49.56 ± 3.84	51.70 ± 3.75	51.56 ± 3.44	52.12 ± 4.16	50.75 ± 4.39	52.04 ± 4.05	-	50.91 ± 3.54	$\mathbf{58.88 \pm 0.0}$
Reuters	56.26 ± 11.67	50.96 ± 9.15	53.79 ± 5.48	54.17 ± 5.27	53.72 ± 5.47	51.25 ± 9.49	60.22 ± 8.62	54.51 ± 10.00	52.18 ± 8.21	$\mathbf{76.06 \pm 0.0}$
COIL20	54.33 ± 4.95	42.77 ± 5.86	58.73 ± 4.56	59.10 ± 4.47	55.00 ± 4.74	55.10 ± 4.42	60.21 ± 4.22	57.87 ± 5.13	58.95 ± 4.67	$\mathbf{66.40 \pm 0.0}$
Letter	35.77 ± 1.22	33.40 ± 1.47	36.95 ± 1.36	33.68 ± 1.53	37.34 ± 1.05	35.54 ± 1.23	36.11 ± 1.32	27.42 ± 0.91	35.40 ± 1.39	$\mathbf{47.15 \pm 0.0}$
NMI(%)										
Mnist	47.82 ± 1.42	45.48 ± 1.53	46.07 ± 1.59	46.55 ± 1.36	46.57 ± 1.27	46.06 ± 1.31	47.62 ± 1.19	-	49.62 ± 1.07	$\mathbf{76.84 \pm 0.0}$
Usps	58.51 ± 3.92	55.89 ± 3.24	57.56 ± 3.63	58.02 ± 3.61	56.21 ± 3.38	57.54 ± 1.51	58.32 ± 1.25	44.60 ± 4.38	62.33 ± 1.37	$\mathbf{74.48 \pm 0.0}$
Fmnist	49.83 ± 3.87	47.03 ± 3.84	49.98 ± 3.75	50.07 ± 3.44	50.19 ± 4.16	50.55 ± 1.16	49.93 ± 1.08	-	49.49 ± 1.03	$\mathbf{61.28 \pm 0.0}$
Reuters	26.37 ± 11.67	21.41 ± 9.15	26.43 ± 5.48	27.37 ± 5.27	25.82 ± 5.47	21.26 ± 9.94	31.99 ± 8.04	23.71 ± 9.67	22.28 ± 11.42	$\mathbf{44.92 \pm 0.0}$
COIL20	68.89 ± 4.95	55.75 ± 5.86	73.38 ± 4.56	74.05 ± 4.47	70.22 ± 4.74	69.14 ± 2.67	74.36 ± 2.38	71.72 ± 2.64	73.43 ± 2.10	$\mathbf{77.72 \pm 0.0}$
Letter	37.58 ± 1.22	34.79 ± 1.47	39.18 ± 1.36	35.06 ± 1.53	39.43 ± 1.05	37.98 ± 0.57	$38.56 \pm 0,57$	30.04 ± 0.63	37.45 ± 0.72	$\mathbf{52.29 \pm 0.0}$
Purity(%)										
Mnist	58.37 ± 1.62	56.64 ± 2.21	57.35 ± 1.79	57.96 ± 1.83	58.09 ± 1.81	56.78 ± 1.89	57.91 ± 1.57	-	59.28 ± 1.47	$\mathbf{84.31 \pm 0.0}$
Usps	69.40 ± 2.47	67.74 ± 2.74	69.11 ± 2.47	69.55 ± 2.00	67.70 ± 2.29	67.72 ± 2.34	69.36 ± 2.45	55.40 ± 5.35	71.35 ± 2.77	$\mathbf{80.19 \pm 0.0}$
Fmnist	56.09 ± 2.20	53.03 ± 2.84	56.71 ± 2.22	56.61 ± 2.42	57.15 ± 1.52	55.98 ± 2.16	56.62 ± 1.87	-	56.27 ± 1.88	$\mathbf{63.47 \pm 0.0}$
Reuters	74.71 ± 5.10	74.21 ± 4.97	78.41 ± 4.08	$\mathbf{78.97 \pm 3.59}$	77.79 ± 3.91	73.78 ± 5.12	78.18 ± 3.94	74.43 ± 4.99	73.67 ± 4.76	75.72 ± 0.0
COIL20	58.03 ± 5.49	45.63 ± 4.97	63.25 ± 3.93	63.96 ± 4.09	59.38 ± 3.83	58.77 ± 4.19	64.63 ± 4.16	61.30 ± 4.04	63.48 ± 3.67	$\mathbf{70.94 \pm 0.0}$
Letter	37.90 ± 1.01	35.23 ± 1.89	39.30 ± 1.02	35.49 ± 1.67	39.76 ± 0.99	38.12 ± 0.99	38.68 ± 0.88	29.04 ± 0.83	37.66 ± 1.14	$\mathbf{49.56 \pm 0.0}$

Table 2: Benchmark Dataset Description

Dataset	Samples	Dimensions	Classes
Mnist	70000	784	10
USPS	9298	256	10
FMNIST	70000	784	10
Reuters-10k	10000	2000	4
COIL-20	1440	1024	20
Letter	20800	784	26

ing dimensions.(2) **Mean-Filling (MF)**: The missing features are imputed with zeros in the normalized data matrix.(3) **Low-rank Completion(LRC)**[20]: The method attempts to recover data matrix with low-rank assumption. (4) **Max Norm Completion (MNC)**[3]: MNC adopts the max-norm to complete missing features.(5)**Factor Group-Sparse Regularization for Efficient Low-Rank Matrix Recovery(FSGR)**[3][3] The author proposes factor group-sparse regularizers to accomplish low-rank matrix completion task.(6)**GAIN**[4][31]: **Missing data imputing using Generative Adversarial Nets**. It proposes a method that uses GAN to estimate and complete the work of filling missing values. (7) **VAEAC:**[5][7]**Variational Autoencoder with Arbitrary Conditioning**. It is a latent variable model trained using stochastic gradient variational Bayes.(8)**MIVAE**[6][18]: MIWAE is based on the importance-weighted autoencoder, and maximises a potentially tight lower bound of the log-likelihood of the observed data. (9)**MDIOT**[7][19]: **Missing Data Imputing using Optimal Transport.** This paper leverages OT to define a loss function for missing data distribution and complete data distribution. The hyper-parameters used in all our comparative experiments follow their corresponding papers.

For all the compared methods above, we have downloaded their public implementations with Matlab and Pytorch. All our experiments are conducted on desktop computer with Intel i7-9700K CPU @ 3.60GHz×12, 64 GB RAM and GeForce RTX 3090 25GB.

4.3. Results Comparisons to Alternative Methods

Table 1 shows the aggregated clustering comparison of the above algorithms on the benchmark datasets. The best results are highlighted with boldface and '-' means the out of GPU memory failure. Based on the results, we have the following observations: **(1)** Our proposed method outperforms all the SOTA imputation competitors in clustering performance by large margins. For example, our algorithm surpasses the second best by **50.4%, 17%, 12%, 26%, 10%** and **30%**, in terms of ACC on all benchmark datasets. In particular, the margins for the four datasets (Mnist, Usps, Reuters and Letter) are very impressive. These results clearly verify the effectiveness of the proposed network. **(2)** Comparing with the generative-style methods, the proposed DDIC-OT consistently further improves the clustering performance and achieves better results among the benchmark datasets. GAIN, VAEAC and

[3]https://github.com/udellgroup/Codes-of-FGSR-for-effecient-low-rank-matrix-recovery
[4]https://github.com/jsyoon0823/GAIN
[5]https://github.com/tigvarts/vaeac
[6]https://github.com/pamattei/miwae
[7]https://github.com/BorisMuzellec/MissingDataOT

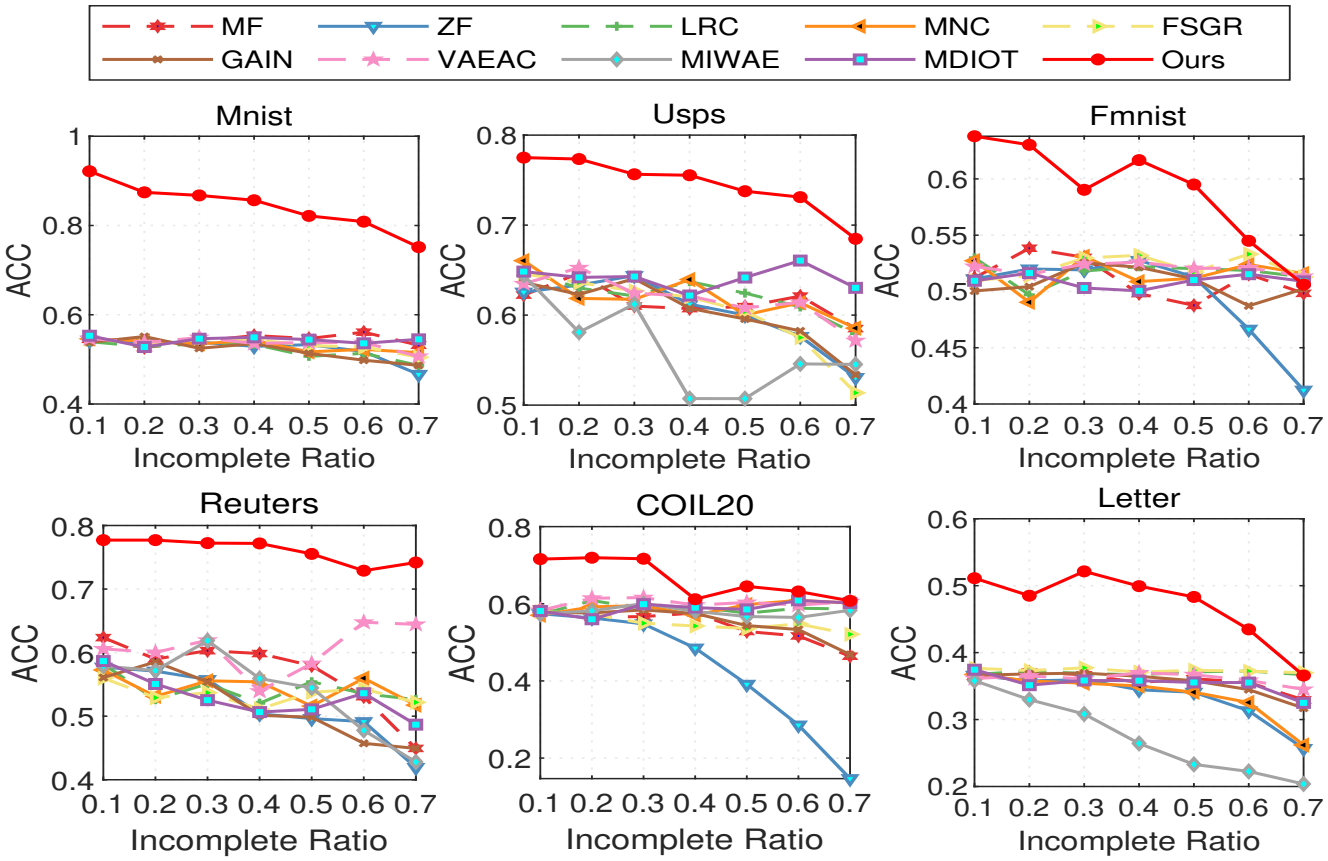


Figure 3: The clustering results of ACC metric on the benchmark datasets with different integrity ratios. The results of Purity are provided in supplementary materials due to space limit.

MIWAE are the chosen representative methods. As can be seen, they concentrate on the generation or imputation task while ignoring the impacts of downstream clustering procedure. The joint optimization framework further contributes to improving performance. **(3)** MDIOT has been considered as a strong baseline for incomplete data imputation. It outperforms other competitors among most of the datasets. Our proposed algorithm surpasses MDIOT by **55.2%, 17.1%, 15.7%, 45.8%, 12.6%** and **33.2%** in terms of ACC on all benchmark datasets. The phenomenon demonstrates the effectiveness of the our proposed architecture. Regardless of directly computing distribution distances in original space, the bottleneck of our embedding layer serves for clustering task and make distributions more discriminate. More details of experimental results are provided in the supplementary materials due to the space limit.

4.4. Qualitative Study

In this section, we deeply analyze the clustering performance regarding to various ratios and the evolution of the learned representation. In order to show the comparison between different methods more clearly, we draw the ACC and NMI of compared methods under different missing rates as line graphs as shown in Figure. 3 and 4.

From the figure, we can obtain the following observations:**(1)** As can be seen, with the incomplete ratios increasing, all the methods suffers the degradation of clustering performance due to more unavailable information. Especially for the generative-based methods (VAEAC and MDIOT), their performance drops sharply due to inaccurate imputations. **(2)** The results of our proposed method in terms of ACC are higher than all the competing algorithms for different integrity ratios. Moreover, our method achieves stable performance against the increasing incomplete ratios. These results clearly demonstrates the effectiveness of DDIC-OT. **(3)** We also show the relative NMI performance of the compared methods in Figure. 4. As can be seen, the clustering performance results are consistent with the ACC observations.

4.5. Ablation Study

Loss Ablation Study

We first investigate how the clustering loss and the distribution-preserving loss affect the clustering perfor-

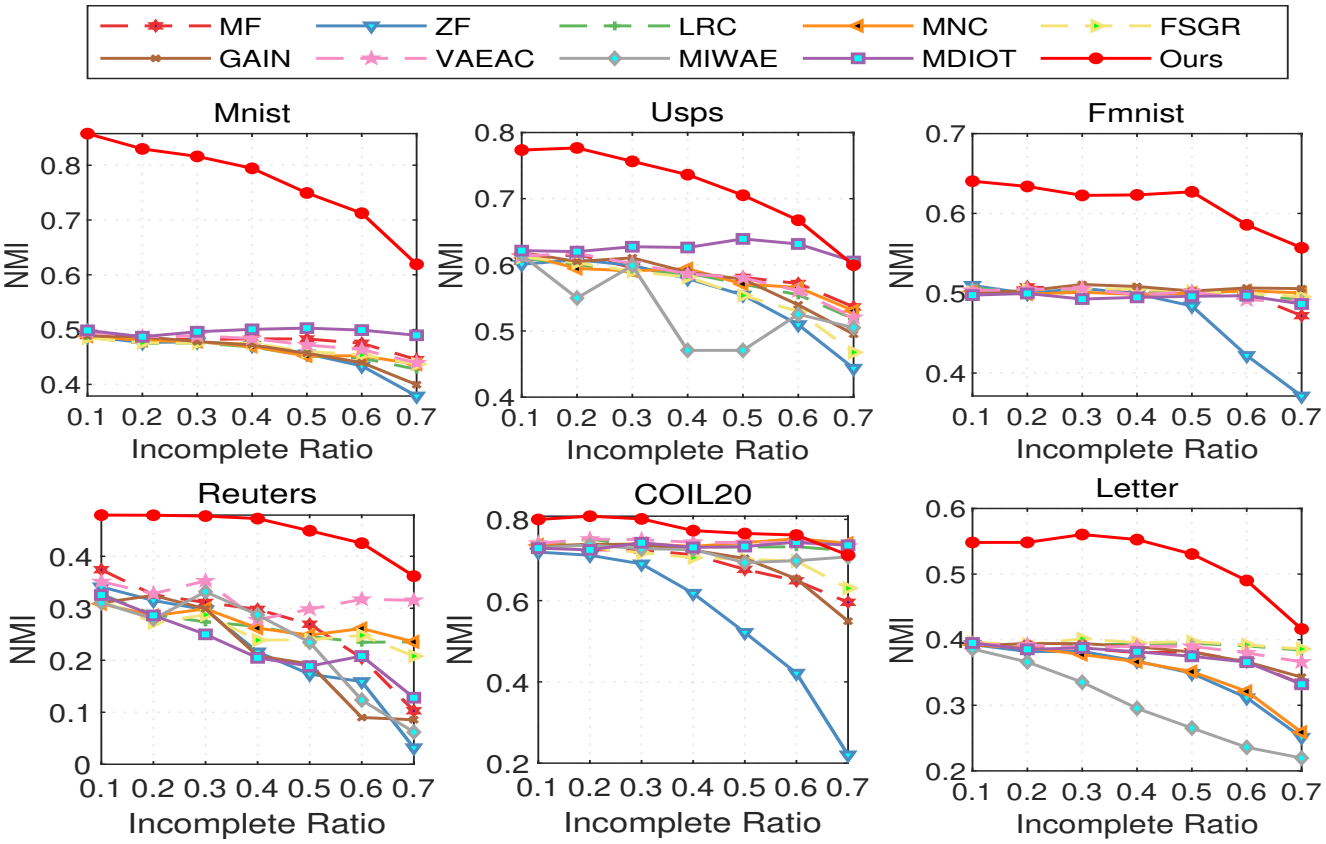


Figure 4: The clustering results of NMI metric on the benchmark datasets with different integrity ratios.

mance on Mnist/Usps/Reuters, and the results are shown in Table 3. In this experiment, we uniformly adopt datasets with 10% missing ratio. It seems that the L_c has more contributions than L_s on Mnist/Usps for clustering, and inversely on Reuters. We also conclude that the joint of two counterpart losses further contributes to better performance.

Table 3: Loss ablation study with 10% missing ratio.

Dataset	Mnist		Usps		Reuters	
Loss	ACC	NMI	ACC	NMI	ACC	NMI
L_s	72.24	62.36	55.59	53.29	76.47	44.58
L_c	86.84	77.7	74.81	73.88	72.42	43.09
$L_s + L_c$	**92.16**	**85.77**	**77.5**	**77.35**	**77.7**	**47.93**

Sensitivity to initialization imputed values

The initialization of imputed values has been demonstrated to be an essential part of incomplete clustering. We tested its sensitivity in our DDIC-OT, w.r.t. model performance on Mnist/Usps/Reuters. We evaluated two commonly-used initialization values: zero-filling (ZF) and mean-filling (MF). Table 4 shows that DDIC-OT can work stably without clear variation in the overall performance when using different initializations. This verifies that our method is insensitive to network initialization.

Table 4: Model sensitivity to different Initialization of imputed values on three benchmarks. Metric:ACC.

Dataset	Mnist		Usps		Reuters	
Loss	ZF	MF	ZF	MF	ZF	MF
10%	90.75	92.16	74.19	77.5	74.03	77.7
30%	83.36	86.74	72.46	75.65	72.47	77.23
50%	78.61	82.15	71.77	73.77	72.03	75.53

5. Conclusion

In this paper, we propose a novel incomplete clustering methods termed DDIC-OT, which jointly performs clustering and missing data imputation into a unified framework. Extensive experiments are conducted to demonstrate the effectiveness of optimal transport for clustering tasks. In the future, we will consider to construct more advanced network to further improve incomplete clustering performance.

References

[1] Nicholas L Crookston and Andrew O Finley. yaimpute: an r package for knn imputation. *Journal of Statistical Software. 23 (10). 16 p.*, 2008. 3

[2] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 3

[3] Jicong Fan, Lijun Ding, Yudong Chen, and Madeleine Udell. Factor group-sparse regularization for efficient low-rank matrix recovery. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019. 2, 6

[4] Zoubin Ghahramani and Michael I Jordan. Supervised learning from incomplete data via an em approach. In *Advances in neural information processing systems*, pages 120–127, 1994. 3

[5] Trisha Greenhalgh, Manuel B Schmid, Thomas Czypionka, Dirk Bassler, and Laurence Gruer. Face masks for the public during the covid-19 crisis. *Bmj*, 369, 2020. 1

[6] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994. 5

[7] O Ivanov, M Figurnov, and D Vetrov. Variational autoencoder with arbitrary conditioning. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. 3, 6

[8] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013. 2

[9] 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

[10] Diederik P Kingma and Max Welling. Stochastic gradient vb and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*, volume 19, 2014. 3

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[13] 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

[14] 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. 1

[15] Xinwang Liu, Xinzhong Zhu, Miaomiao Li, Lei Wang, Chang Tang, Jianping Yin, Dinggang Shen, Huaimin Wang, and Wen Gao. Late fusion incomplete multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2410–2423, 2018. 1

[16] 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

[17] Si Lu, Xiaofeng Ren, and Feng Liu. Depth enhancement via low-rank matrix completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3390–3397, 2014. 2

[18] Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, pages 4413–4423. PMLR, 2019. 6

[19] Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pages 7130–7140. PMLR, 2020. 4, 6

[20] Feiping Nie, Heng Huang, and Chris Ding. Low-rank matrix recovery via efficient schatten p-norm minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 2012. 2, 6

[21] 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. 1

[22] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 3

[23] Trevor W. Richardson, Wencheng Wu, Lei Lin, Beilei Xu, and Edgar A. Bernal. Mcflow: Monte carlo flow models for data imputation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[24] Yuandong Tian, Wei Liu, Rong Xiao, Fang Wen, and Xiaoou Tang. A face annotation framework with partial clustering and interactive labeling. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1

[25] 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. 1

[26] Zaiwen Wen, Wotao Yin, and Yin Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012. 2

[27] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8150–8159, 2019. 1

[28] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International*

conference on machine learning, pages 478–487. PMLR, 2016. 4

[29] Congyuan Yang, Daniel Robinson, and Rene Vidal. Sparse subspace clustering with missing entries. In *International Conference on Machine Learning*, pages 2463–2472. PMLR, 2015. 1

[30] Liu Yang, Chenyang Shen, Qinghua Hu, Liping Jing, and Yingbo Li. Adaptive sample-level graph combination for partial multiview clustering. *IEEE Transactions on Image Processing*, 29:2780–2794, 2019. 1

[31] Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, pages 5689–5698. PMLR, 2018. 2, 3, 6

[32] 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. 1

[33] Changqing Zhang, Huazhu Fu, Si Liu, Guangcan Liu, and Xiaochun Cao. Low-rank tensor constrained multiview subspace clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 1582–1590, 2015. 1